Joint Channel Estimation and Data Detection for Multihop OFDM Relaying System under Unknown Channel Orders and Doppler Frequencies

Rui Min and Yik-Chung Wu*

Abstract

In this paper, channel estimation and data detection for multihop relaying orthogonal frequency division multiplexing (OFDM) system is investigated under time-varying channel. Different from previous works, which highly depend on the statistical information of the doubly-selective channel (DSC) and noise to deliver accurate channel estimation and data detection results, we focus on more practical scenarios with unknown channel orders and Doppler frequencies. Firstly, we integrate the multilink, multihop channel matrices into one composite channel matrix. Then, we formulate the unknown channel using generalized complex exponential basis expansion model (GCE-BEM) with a large oversampling factor to introduce channel sparsity on delay-Doppler domain. To enable the identification of nonzero entries, sparsity enhancing Gaussian distributions with Gamma hyperpriors are adopted. An iterative algorithm is developed under variational inference (VI) framework. The proposed algorithm iteratively estimate the channel, recover the unknown data using Viterbi algorithm and learn the channel and noise statistical information, using only limited number of pilot subcarrier in one OFDM symbol. Simulation results show that, without any statistical information, the performance of the proposed algorithm is very close to that of the optimal channel estimation and data detection algorithm, which requires specific information on system structure, channel tap positions, channel lengths, Doppler shifts as well as noise powers.

Rui Min and Yik-Chung Wu are with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong. Email:{minrui,ycwu}@eee.hku.hk.

*The corresponding author is Yik-Chung Wu.

Index Terms

Doubly-selective channel, Channel estimation, Data detection, Variational inference, Orthogonal frequency division multiplexing, Multihop relaying system

I. Introduction

Next generation broadband system aims to support higher levels of mobility, connectivity and efficiency. Multihop relaying system is a perfect suit for such requirement due to their benefits in easy deployment, enhanced connectivity, flexible adaptability, and increased capacity. On the other hand, orthogonal frequency division multiplexing (OFDM) has been adopted as the transmission scheme for many next generation broadband standards, such as WiMAX, LTE and IEEE 802.16. These result in the need to develop receiver algorithms for multihop OFDM system under high mobility. With high mobility, the broadband wireless channel is both frequency-selective and time-varying, a.k.a. doubly-selective. The channel responses vary sample by sample, which destroy the orthogonal property among subcarriers and causes intercarrier interference (ICI). Besides, the relaying system structure and channel statistical information are generally unknown to the receiver, due to flexible configuration of relaying paths. These poses strong challenges to channel estimation and data detection of OFDM relaying system under high mobility.

Over doubly-selective channel (DSC), channel estimation and data detection for single-hop OFDM systems has been considered in [1]–[6], where the two tasks are treated separately. In [1]–[4], the frequency domain channel matrix is approximated with a diagonal matrix under the assumption of small normalized Doppler frequency. The resulting algorithm would produce poor channel estimate and subsequently degrade the data detection performance for fast time-varying DSC. In view of that, [5] and [6] assumed a banded frequency-domain channel matrix, thus achieve a better channel modeling accuracy. However, due to the ICI introduced in frequency domain, pilots and data would interfere each other. This leads to the interdependence between channel estimation and data detection, and joint processing of them is necessary.

Research on multihop channel estimation is still limited especially for DSC, due to the complicated system structure and uncertain time-varying channel property. Among the limited existing works, [7] studied the two-way relaying (TWR) system under frequency-nonselective time-varying channel, and complex exponential basis expansion model (CE-BEM) was used to reduce the number of channel parameters. However, [7] only considered single carrier system and extension to multicarrier system is not straightforward due to additional ICI. More recently, in [8], an iterative algorithm for data detection and channel estimation was proposed for dual-hop amplify-and-forward (AF) OFDM system. With complete information of channel and noise in each hop, the data detection results were very close to the ideal case. Unfortunately, this work did not consider multihop relaying system, which has a more complicated structure. More importantly, it requires the destination receiver to have full statistical information of all channels and noise powers, which might not be readily available in practice.

In this paper, we study the channel estimation and data detection of OFDM-based multihop AF relaying system under high mobility, with special focus on unknown channel orders and Doppler frequencies. Based on the fact that the combined channel information is sufficient for data detection, and in order to reduce computation load on relays and time delay of the whole system, no channel estimation is performed at the relays. Different from previous works which highly rely on information of system structure, channel tap positions, channel lengths and Doppler frequencies of all channels, as well as noise powers at all receivers, we propose to solve the problems with none of the above information. By first expanding the composite source-relay-destination channel using generalized complex exponential basis expansion model (GCE-BEM) with a large oversampling factor, we introduce channel sparsity on delay-Doppler domain. Then sparsity enhancing Gaussian distributions with Gamma hyperpriors are adopted for channel estimation to enable the identification of nonzero elements. An iterative algorithm is proposed based on variational inference (VI) framework to iteratively estimate the channel, recover the unknown data using Viterbi algorithm and learn the channel and noise statistical information, using only limited number of pilot subcarrier in one OFDM symbol. Simulation results show that the performance of the proposed algorithm is very close to that of an optimal algorithm, which requires detailed statistical information on channels and noises.

The rest of the paper is organized as follows. The OFDM-based multihop relaying system is introduced in Section II. Then in Section III, the channel matrices of all the hops are integrated into one concise composite channel matrix and reformulated using GCE-BEM with a large oversampling factor. The iterative channel estimation and data detection algorithm is developed under VI framework in Section IV. And in Section V, least squares (LS) channel estimator and equalizer are derived to obtain the initial parameters for the proposed iterative algorithm.

Simulation results of the proposed algorithm are presented in Section VI. Finally, this paper is concluded in Section VII.

Notations: Boldface uppercase and lowercase letters will be used for matrices and vectors, respectively. Superscripts H, T and $*$ denote Hermitian, transpose and conjugate, respectively. The symbol $\mathbf{I}_N$ represents the $N \times N$ identity matrix. Symbol $\mathbf{e}_l$ denotes the vector with structure given as $\left[\mathbf{0}_{1\times l}, 1, \mathbf{0}_{1\times(N-l-1)}\right]^T$, where $\mathbf{0}_{1\times l}$ is the l dimension all-zero row vector. $\mathrm{diag}\{\mathbf{x}\}$ stands for the diagonal matrix with vector $\mathbf{x}$ on its diagonal. The notation $[\mathbf{X}]_{m_1:m_2,n_1:n_2}$ represents the submatrix of $\mathbf{X}$ consists of entries on the m_1-to-m_2^{th} rows and n_1-to-n_2^{th} columns. $\mathbb{E}\{\cdot\}$ denotes the expectation while $\mathrm{Tr}\{\mathbf{X}\}$ and $\det\{\mathbf{X}\}$ are the trace and the determinant of the square matrix $\mathbf{X}$. $\mathrm{Re}\{\cdot\}$ denotes the real part. And $\lceil x \rceil$ rounds x to the nearest integer greater than or equal to x. Finally, $\mathbf{F}$ represents the discrete Fourier transform (DFT) matrix with $[\mathbf{F}]_{m,n} = \frac{1}{\sqrt{N}} e^{-j2\pi mn/N}$.

II. System Model

In this paper, we consider a multihop relaying system consists of a source $\mathbb{S}$, a destination $\mathbb{D}$ and a number of relays scattered in the middle. Each of them is equipped with single antenna. Without loss of generality, we assume the relays work cooperatively to form K links, each of them consisting of $\Upsilon + 1$ hops. Apart from the K relaying paths, there is no other link between $\mathbb{S}$ and $\mathbb{D}$, and all relays employ the AF scheme. Denoting the relay on the k^{th} link connecting the ρ^{th} and the $(\rho+1)^{th}$ hop as $\mathbb{R}_{k,\rho}$, the relaying system is shown in Figure 1.

The channel of each hop is assumed to be doubly-selective channel (DSC). Specifically, at the ρ^{th} hop of the k^{th} relaying path, the channel consists of $N_{k,\rho}$ independent nonzero channel taps with maximum delay of $(L_{\max}^{k,\rho} - 1)T_s$, where T_s is the sample interval. We consider the general situation that the channel taps are not necessarily consecutive, so that we have $N_{k,\rho} \le L_{\max}^{k,\rho}$. Let $\bar{h}_{k,\rho}(n,l)$ be the l^{th} tap of that channel at time nT_s. For a given ρ and k, the channel taps are independent and each one being a zero-mean complex Gaussian process with bandlimited power spectral density within $[-f_{k,\rho}(l), f_{k,\rho}(l)]$, where $f_{k,\rho}(l)$ is the maximum Doppler shift of the l^{th} tap. In general, $f_{k,\rho}(l)$ may be distinct for different l, since each tap results from signal transmission through a different physical scattering. Furthermore, it is assumed that the channels for different links k and hops ρ are independent from each other.

A. OFDM Signal Transmitted from $\mathbb{S}$

In an OFDM system, the frequency domain source data $\mathbf{x} = [x(0), \ldots, x(N-1)]^T$ is first transformed to the time domain data $\mathrm{s} = \mathbf{F}^H\mathbf{x}$, where $\mathbf{F}$ represents the discrete Fourier transform (DFT) matrix. In order to facilitate channel estimation and data detection, pilots are inserted in the frequency domain as

$$x(n) = \begin{cases} x_p(n) & \forall \quad n \in \mathfrak{I}_p \\ x_d(n) & \forall \quad n \in \mathfrak{I}_d, \end{cases} \tag{1}$$

where $\mathfrak{I}_d$ is the index set of the N_d unknown data symbols, $\mathfrak{I}_p$ is the index set of the N_p pilot symbols and we have $N = N_d + N_p$. In matrix form, $\mathbf{x}$ can be represented as

$$\mathbf{x} = \mathbf{E}_d\mathbf{x}_d + \mathbf{E}_p\mathbf{x}_p, \tag{2}$$

where $\mathbf{E}_d$ and $\mathbf{E}_p$, with dimensions $N \times N_d$ and $N \times N_p$, respectively, map $\mathbf{x}_d$ and $\mathbf{x}_p$ to their corresponding subcarriers. Before transmission, a Cyclic Prefix (CP) of length L_{cp} is added at the beginning of s to prevent intersymbol interference (ISI). Since the OFDM signal goes through a number of relays before reaching the destination, L_{cp} should be larger than the maximum channel length among all the relaying paths, denoted as $L_{\max} = \max\limits_k(\sum_{\rho=1}^{\Upsilon+1} L_{\max}^{k,\rho} - \Upsilon)$.

B. Received OFDM Signal

In AF relaying system, each relay merely amplify the received signal before passing the signal to the next relay or destination. For the k^{th} relaying path, the signal received at $\mathbb{R}_{k,1}$ is given by

$$r_{k,1}(n) = \sum_{l=0}^{L_{\max}^{k,1}-1} \bar{h}_{k,1}(n,l)s(n-l) + w_{k,1}(n), \tag{3}$$

where $w_{k,1}(n)$ is the additive white Gaussian noise (AWGN) with power $\varpi_{k,1}^2$. Upon receiption, the relay amplifies the incoming signal as [9]

$$z_{k,1}(n) = \varsigma_{k,1}r_{k,1}(n) \tag{4}$$

and then transmits $z_{k,1}(n)$ to the next relay $\mathbb{R}_{k,2}$ through channel $\bar{h}_{k,2}(n,l)$. The received signal is then represented as

$$r_{k,2}(n) = \sum_{l=0}^{L_{\max}^{k,2}-1} \bar{h}_{k,2}(n,l)z_{k,1}(n-l) + w_{k,2}(n), \tag{5}$$

with $w_{k,2}(n)$ being the AWGN with power $\varpi_{k,2}^2$. The signal is then amplified as $z_{k,2}(n) = \varsigma_{k,2} r_{k,2}(n)$ and transmitted to $\mathbb{R}_{k,3}$ and so on. Then, at the Υ^{th} relay, the amplified signal is transmitted to the destination. Finally at destination $\mathbb{D}$, the received signal is given by

$$\tilde{y}(n) = \sum_{k=1}^{K} \sum_{l=0}^{L_{\max}^{k,\Upsilon+1}-1} \bar{h}_{k,\Upsilon+1}(n,l) z_{k,\Upsilon}(n-l) + w_d(n), \tag{6}$$

where AWGN $w_d(n)$ has power ϖ_d^2. Upon reception, the CP $[\tilde{y}(-L_{cp}), \ldots, \tilde{y}(-1)]^T$ is removed and the received signal $\tilde{\mathbf{y}} = [\tilde{y}(0), \ldots, \tilde{y}(N-1)]^N$ can be written in matrix form as

$$\tilde{\mathbf{y}} = \sum_{k=1}^{K} \bar{\mathbf{H}}_{k,\Upsilon+1} \mathbf{z}_{k,\Upsilon} + \mathbf{w}_d, \tag{7}$$

where $\mathbf{z}_{k,\Upsilon} = [z_{k,\Upsilon}(-(L_{\max}^{k,\Upsilon+1} - 1)), \ldots, z_{k,\Upsilon}(0), \ldots, z_{k,\Upsilon}(N-1)]^T$, $\mathbf{w}_d$ is the noise vector with elements $w_d(n)$, and $\bar{\mathbf{H}}_{k,\Upsilon+1}$ is an $N \times (N + L_{\max}^{k,\Upsilon+1} - 1)$ channel matrix given by

$$\bar{\mathbf{H}}_{k,\Upsilon+1} = \begin{bmatrix} \bar{h}_{k,\Upsilon+1}(0, L_{\max}^{k,\Upsilon+1} - 1) & \ldots & \bar{h}_{k,\Upsilon+1}(0,0) & & \\ & \bar{h}_{k,\Upsilon+1}(1, L_{\max}^{k,\Upsilon+1} - 1) & \ldots & \bar{h}_{k,\Upsilon+1}(1,0) & \\ & & \cdots & & \\ & & \bar{h}_{k,\Upsilon+1}(N-1, L_{\max}^{k,\Upsilon+1} - 1) & \ldots & \bar{h}_{k,\Upsilon+1}(N-1,0) \end{bmatrix}. \tag{8}$$

Furthermore, $\mathbf{z}_{k,\Upsilon}$ can be written in terms of $\mathbf{z}_{k,\Upsilon-1}$ as

$$\mathbf{z}_{k,\Upsilon} = \varsigma_{k,\Upsilon} \bar{\mathbf{H}}_{k,\Upsilon} \mathbf{z}_{k,\Upsilon-1} + \varsigma_{k,\Upsilon} \mathbf{w}_{k,\Upsilon}, \tag{9}$$

with $\mathbf{z}_{k,\Upsilon-1} = [z_{k,\Upsilon-1}(-(L_{\max}^{k,\Upsilon+1} + L_{\max}^{k,\Upsilon}) + 2), \ldots, z_{k,\Upsilon-1}(0), \ldots, z_{k,\Upsilon-1}(N-1)]^T$, $\mathbf{w}_{k,\Upsilon}$ is the corresponding noise vector, and $\bar{\mathbf{H}}_{k,\Upsilon}$ is an $(N + L_{\max}^{k,\Upsilon+1} - 1) \times (N + L_{\max}^{k,\Upsilon+1} + L_{\max}^{k,\Upsilon} - 2)$ matrix given by

$$\bar{\mathbf{H}}_{k,\Upsilon} = \begin{bmatrix} \bar{h}_{k,\Upsilon}(1 - L_{\max}^{k,\Upsilon+1}, L_{\max}^{k,\Upsilon} - 1) & \ldots & \bar{h}_{k,\Upsilon}(1 - L_{\max}^{k,\Upsilon+1}, 0) & & \\ & \bar{h}_{k,\Upsilon}(2 - L_{\max}^{k,\Upsilon+1}, L_{\max}^{k,\Upsilon} - 1) & \ldots & \bar{h}_{k,\Upsilon}(2 - L_{\max}^{k,\Upsilon+1}, 0) & \\ & & \cdots & & \\ & & \bar{h}_{k,\Upsilon}(N-1, L_{\max}^{k,\Upsilon} - 1) & \ldots & \bar{h}_{k,\Upsilon}(N-1,0) \end{bmatrix}. \tag{10}$$

Tracing back to the 1^{st} hop, we have $\mathbf{z}_{k,1} = \varsigma_{k,1} \bar{\mathbf{H}}_{k,1} \mathbf{s}_k + \varsigma_{k,1} \mathbf{w}_{k,1}$, where $\bar{\mathbf{H}}_{k,1}$ is an $(N + \sum_{\rho=2}^{\Upsilon+1} L_{\max}^{k,\rho} - \Upsilon) \times (N + \sum_{\rho=1}^{\Upsilon+1} L_{\max}^{k,\rho} - \Upsilon - 1)$ channel matrix with structure the same as (8) and (10), and $\mathbf{s}_k = \mathbf{E}_k \mathbf{s}$ with $\mathbf{E}_k = [[\mathbf{I}_N]_{1:N,(N - \sum_{\rho=1}^{\Upsilon+1} L_{\max}^{k,\rho} + \Upsilon + 2):N}, \mathbf{I}_N]^T$ characterizing the effect of the CP. Based on the above derivations, the received signal vector $\tilde{\mathbf{y}}$ is

$$\begin{aligned}\tilde{\mathbf{y}} \;&=\; \underbrace{\sum_{k=1}^{K}\left[\left(\prod_{\rho=1}^{\Upsilon}\varsigma_{k,\rho}\right)\left(\bar{\mathbf{H}}_{k,\Upsilon+1}\dots\bar{\mathbf{H}}_{k,1}\right)\mathbf{E}_k\right]}_{\triangleq\mathbf{H}}\mathbf{F}^H\mathbf{x}\\ &\quad\underbrace{+\sum_{k=1}^{K}\left[\sum_{\rho=1}^{\Upsilon}\left(\left(\prod_{\varrho=\rho}^{\Upsilon}\varsigma_{k,\varrho}\right)\left(\bar{\mathbf{H}}_{k,\Upsilon+1}\dots\bar{\mathbf{H}}_{k,\rho+1}\right)\right)\mathbf{w}_{k,\rho}\right]+\mathbf{w}_d,}_{\triangleq\tilde{\mathbf{v}}}\end{aligned} \qquad (11)$$

where $\mathbf{H}$ represents the composite channel matrix and $\tilde{\mathbf{v}}$ represents the composite noise effect.

III. Reformulation of the Composite Channel Matrix

In order to estimate the channel and detect the data, it is important to investigate the structure of the channel matrix $\mathbf{H}$. Writing $\mathbf{H} = \sum_{k=1}^{K}\left(\prod_{\rho=1}^{\Upsilon}\varsigma_{k,\rho}\right)\mathbf{H}_k$, where $\mathbf{H}_k = \bar{\mathbf{H}}_{k,\Upsilon+1}\bar{\mathbf{H}}_{k,\Upsilon}\dots\bar{\mathbf{H}}_{k,1}\mathbf{E}_k$. To find out the structure of $\bar{\mathbf{H}}_{k,\Upsilon+1}\bar{\mathbf{H}}_{k,\Upsilon}\dots\bar{\mathbf{H}}_{k,1}\mathbf{E}_k$, we start from $\bar{\mathbf{H}}_{k,\Upsilon+1}$ and $\bar{\mathbf{H}}_{k,\Upsilon}$ with their expressions given in (8) and (10), respectively. Each matrix represents the linear convolution of a time-varying channel and the matrix multiplication expresses the convolution effect of two time-varying channels. Therefore the resulting matrix $\bar{\mathbf{H}}_{k,\Upsilon+1}\bar{\mathbf{H}}_{k,\Upsilon}$ will also be in the form of (8) and (10), except that the resulting channel length of the new time-varying channel is now being $L_{\max}^{k,\Upsilon+1} + L_{\max}^{k,\Upsilon} - 1$.

Similarly, multiplying $\bar{\mathbf{H}}_{k,\Upsilon-1}$ to $\bar{\mathbf{H}}_{k,\Upsilon+1}\bar{\mathbf{H}}_{k,\Upsilon}$ from the right, the result $\bar{\mathbf{H}}_{k,\Upsilon+1}\bar{\mathbf{H}}_{k,\Upsilon}\bar{\mathbf{H}}_{k,\Upsilon-1}$ will be an $N\times(N+\sum_{\rho=\Upsilon-1}^{\Upsilon+1}-3)$ matrix, with equivalent channel length of $\sum_{\rho=\Upsilon-1}^{\Upsilon+1}L_{\max}^{k,\rho}-2$, due to the convolution effect. Continuing the matrix multiplication, we have $\bar{\mathbf{H}}_{k,\Upsilon+1}\dots\bar{\mathbf{H}}_{k,1}$ being an $N\times(N+\sum_{\rho=1}^{\Upsilon+1}L_{\max}^{k,\rho}-\Upsilon-1)$ matrix, with equivalent channel length of $\sum_{\rho=1}^{\Upsilon+1}L_{\max}^{k,\rho}-\Upsilon$. And eventually $\mathbf{E}_k$ moves the $\sum_{\rho=1}^{\Upsilon+1}L_{\max}^{k,\rho}-\Upsilon-1$ columns from the left part of $\bar{\mathbf{H}}_{k,\Upsilon+1}\dots\bar{\mathbf{H}}_{k,1}$ to the upper right corner. The resulted composite channel matrix $\mathbf{H}_k$ is an $N\times N$ circular convolution matrix of a time-varying channel with equivalent channel length of $\sum_{\rho=1}^{\Upsilon+1}L_{\max}^{k,\rho}-\Upsilon$. Thus $\mathbf{H}$, as the weighted sum of $\mathbf{H}_k$'s, has the same circular convolution matrix structure of a time-varying channel with length $L_{\max} = \max_{k}(\sum_{\rho=1}^{\Upsilon+1}L_{\max}^{k,\rho}-\Upsilon)$:

$$\mathbf{H} = \left[\begin{array}{ccccc} \bar{\mu}(0,0) & \mathbf{0} & \bar{\mu}(0,L_{\max}-1) & \dots & \bar{\mu}(0,1)\\ \bar{\mu}(1,1) & \bar{\mu}(1,0) & \mathbf{0} & \bar{\mu}(1,L_{\max}-1) & \dots\\ & \dots & \dots & \dots & \\ & \mathbf{0} & \bar{\mu}(N-1,L_{\max}-1) & \dots & \bar{\mu}(N-1,0) \end{array}\right], \qquad (12)$$

or equivalently

$$\mathbf{H} = \sum_{l=0}^{L_{\max}-1} \mathrm{diag}\{\bar{\boldsymbol{\mu}}_l\}\mathbf{P}(l), \tag{13}$$

where $\bar{\boldsymbol{\mu}}_l = [\bar{\mu}(0,l), \ldots, \bar{\mu}(N-1,l)]^T$ consists of all the composite channel coefficients of the l^{th} tap and $\mathbf{P}(l) = [\mathbf{e}_l, \ldots, \mathbf{e}_{N-1}, \mathbf{e}_0, \ldots, \mathbf{e}_{l-1}]$. Thus (11) becomes

$$\tilde{\mathbf{y}} = \sum_{l=0}^{L_{\max}-1} \mathrm{diag}\{\bar{\boldsymbol{\mu}}_l\}\mathbf{P}(l)\mathbf{F}^H\mathbf{x} + \tilde{\mathbf{v}}. \tag{14}$$

It should be noticed that, the receiver knows neither the individual channel information of each hop nor the statistical information about the composite channel. This is a natural assumption, as the channels are time-varying and depend on the speed of transceivers and the environment around them. In particular, the receiver has no knowledge on the composite channel tap positions (if the tap positions of individual channel are not consecutive) and the maximum Doppler shift $f_{\max} = \max_k \left(\sum_{\rho=1}^{\Upsilon+1} \max_{l\in[0,L_{\max}^{k,\rho}]} f_{k,\rho}(l)\right)$. Furthermore, the noise power at each relay $\mathbb{R}_{k,\rho}$ is not available to the receiver either. As a result, the receiver has no information on the composite noise power.

In order to proceed, we propose to calculate an upper bound on the maximum Doppler shift and the delay for the composite channel. Let $v_{\max}$ be the maximum relative velocity between two units in any hop in the relaying system. Since $v_{\max}f_c/c \geq f_{k,\rho}(l)$ for all k, ρ and l, where f_c and c are the carrier frequency and the speed of light, respectively, we have $f_{\max} \leq f_U = (\Upsilon+1)v_{\max}f_c/c$. And in the delay domain, the best the receiver knows is that L_{cp} is chosen large enough to avoid ISI. Thus $L_{\max} \leq L_{cp}$ and all the nonzero taps fall in the range of $\{0, \ldots, L_{cp}-1\}$. With the ranges of the delay-Doppler domain defined for the composite channel, we can expand the channel with generalized complex exponential basis expansion model (GCE-BEM) as follows

$$\bar{\mu}(n,l) = \sum_{q=-Q}^{Q} \mu_q(l)e^{j2\pi qn/VN}, \quad l = 0, \ldots, L_{cp}-1, \quad n = 0, \ldots, N-1, \tag{15}$$

where $Q = \lceil VNf_UT_s \rceil$ and V is the oversampling factor, and $\mu_q(l)$ is the GCE-BEM coefficient. It should be noticed that $\mu_q(l) = 0$ in two conditions: 1) $\bar{\mu}(n,l) = 0$; 2) $|q| > VNf_{\max}T_s$.

From (15), the vector $\bar{\boldsymbol{\mu}}_l$ can be expressed as $\bar{\boldsymbol{\mu}}_l = \sum_{q=-Q}^{Q} \boldsymbol{\varphi}(q)\mu_q(l)$, where $\boldsymbol{\varphi}(q) = [1, e^{j2\pi q/VN}, \ldots, e^{j2\pi q(N-1)/VN}]^T$ denotes the q^{th} basis vector. Putting this result into (14), taking the DFT on the signal $\tilde{\mathbf{y}}$ and replacing the unknown $L_{\max}$ with L_{cp}, we have

$$\begin{aligned}\mathbf{y} = \mathbf{F}\tilde{\mathbf{y}} &= \sum_{l=0}^{L_{cp}-1} \mathbf{F}\mathrm{diag}\{\sum_{q=-Q}^{Q} \boldsymbol{\varphi}(q)\mu_q(l)\}\mathbf{P}(l)\mathbf{F}^H\mathbf{x} + \mathbf{v} \\ &= \sum_{l=0}^{L_{cp}-1} \sum_{q=-Q}^{Q} [\mathbf{F}\mathrm{diag}\{\boldsymbol{\varphi}(q)\}\mathbf{P}(l)\mathbf{F}^H\mathbf{x}]\mu_q(l) + \mathbf{v}, \end{aligned} \tag{16}$$

where $\mathbf{v} = \mathbf{F}\tilde{\mathbf{v}}$ represents the noise vector after DFT. Let $\boldsymbol{\mu}_q = [\mu_q(0), \ldots, \mu_q(L_{cp}-1)]^T$, then (16) can be written as

$$\mathbf{y} = \sum_{q=-Q}^{Q} \underbrace{[\mathbf{F}\mathrm{diag}\{\boldsymbol{\varphi}(q)\}\mathbf{P}(0)\mathbf{F}^H\mathbf{x}, \ldots, \mathbf{F}\mathrm{diag}\{\boldsymbol{\varphi}(q)\}\mathbf{P}(L_{cp}-1)\mathbf{F}^H\mathbf{x}]}_{\triangleq \mathbf{G}_q[\mathbf{x}]}\boldsymbol{\mu}_q + \mathbf{v}. \tag{17}$$

Further define $\boldsymbol{\mu} = [\boldsymbol{\mu}_{-Q}^T, \ldots, \boldsymbol{\mu}_Q^T]^T$ and let $\mathbf{G}[\mathbf{x}] = [\mathbf{G}_{-Q}[\mathbf{x}], \ldots, \mathbf{G}_Q[\mathbf{x}]]$, thus we have

$$\mathbf{y} = \mathbf{G}[\mathbf{x}]\boldsymbol{\mu} + \mathbf{v}. \tag{18}$$

On the other hand, from (11), let $\mathbf{D}[\boldsymbol{\mu}] = \mathbf{F}\mathbf{H}\mathbf{F}^H$, the system model can also be written as

$$\mathbf{y} = \mathbf{D}[\boldsymbol{\mu}]\mathbf{x} + \mathbf{v}. \tag{19}$$

It is clear that $\mathbf{D}[\boldsymbol{\mu}]\mathbf{x} = \mathbf{G}[\mathbf{x}]\boldsymbol{\mu}$.

IV. Iterative Channel Estimation and Data Detection

From the system model (18) and (19), the problem is to jointly estimate the composite channel BEM coefficients $\boldsymbol{\mu}$ and the unknown data $\mathbf{x}_d$, without the knowledge of the composite noise variance, denoted by ϖ_v^2. Since we have expanded the composite channel over an extended range in the delay-Doppler plane, we also want to make use of the prior information that most of the BEM coefficients will be zero (i.e., $\boldsymbol{\mu}$ is sparse). It is noticed from (18) and (19) that, estimation of channel requires knowledge of data and vice versa, thus leads to challenges in joint channel estimation and data detection. In this paper, a variational framework is adopted to iteratively improve the channel estimation and data detection results. Compared with other iterative frameworks, e.g., expectation-maximization (EM), VI is more general as it works within a complete Bayesian paradigm and gives a posterior distribution over all the parameters. Below, we first assign prior distributions to the unknown parameters.

A. Prior Distributions of the Unknown Parameters

First, the prior distribution of $\boldsymbol{\mu}$ is assumed to be Gaussian

$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{1}{\pi^M \det(\mathbf{A}^{-1})} \exp\{\boldsymbol{\mu}^H \mathbf{A} \boldsymbol{\mu}\}, \tag{20}$$

where $M = (2Q+1)L_{cp}$, $\mathbf{A} = \mathrm{diag}\{\boldsymbol{\alpha}\}$ and $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_M]^T$ is a vector containing the inverse variance of the elements of $\boldsymbol{\mu}$. Then a hyperprior for $\boldsymbol{\alpha}$ is specified as [10]

$$p(\alpha_j) = \mathrm{Ga}(\alpha_j|a_j, b_j) = b_j^{a_j} \alpha_j^{a_j - 1} \exp(-b_j \alpha_j)/\Gamma(a_j), \tag{21}$$

with parameters a_j, b_j. Although the Gaussian prior given by (20) does not have strong probability peaks for sparsity promotion, by working with (21), the marginal prior $p(\boldsymbol{\mu})$ obtained by integrating out $\boldsymbol{\alpha}$ is a $t-$distribution, which nicely approximates a Laplace distribution [10]. Laplace prior is widely adopted in $L1$-norm regularization schemes like Basis Pursuit (BP) [11]. Unfortunately, using the Laplace prior directly does not lead to a tractable variational treatment [10]. As a result, BP is usually used in one-shot sparse channel estimation [12]–[14]. Furthermore, BP or BP denoising methods rely on the noise power information [15], which is not known in our case. Thus, the above hierarchical prior structure, which is both sparsity promoting and analytically tractable, is a suitable alternative for our problem.

For $\mathbf{x}_d$, since we do not have knowledge on its value before observing the received signal, we set equal preference to all constellation points. Furthermore, due to the independent property among data elements, we have

$$p(\mathbf{x}_d) = \frac{1}{\mathcal{M}_d^{N_d}} \prod_{n=1}^{N_d} \Big[\sum_{\bar{x}_d(n) \in \mathbb{C}_d} \delta(x_d(n) - \bar{x}_d(n)) \Big], \tag{22}$$

where $\mathbb{C}_d$ is the constellation points of the modulation and $\mathcal{M}_d$ is the modulation order, e.g., $\mathcal{M}_d = 4$ for QPSK.

Besides, the unknown noise power is assumed to obey a Gamma prior, such that it can be learned under the variational framework. For ease of expression, let $\beta = 1/\varpi_v^2$ and then

$$p(\beta) = \mathrm{Ga}(\beta|c, d) = d^c \beta^{c-1} \exp(-d\beta)/\Gamma(c), \tag{23}$$

where c, d are the parameters of the Gamma distribution. In the absence of prior information, small values for hyperparameters are chosen, i.e., $a_j = b_j = c = d = 10^{-6}$, so as to produce uninformative priors for the channel and noise power [10].

B. Variational Inference

With the introduced prior and hyperprior distributions, our aim is to jointly estimate $\boldsymbol{\mu}, \boldsymbol{\alpha}, \beta$ and $\mathbf{x}_d$. In Bayesian framework, this corresponds to maximizing the posterior probability density function (pdf) $p(\boldsymbol{\mu}, \boldsymbol{\alpha}, \beta, \mathbf{x}_d|\mathbf{y})$. However, this pdf is in general very hard to be obtained in closed-form and the maximization of it is inconvenient. In the VI framework, a $Q(\boldsymbol{\mu}, \boldsymbol{\alpha}, \beta, \mathbf{x}_d)$ function, which is in tractable form but closely represents $p(\boldsymbol{\mu}, \boldsymbol{\alpha}, \beta, \mathbf{x}_d|\mathbf{y})$, is adopted to efficiently derive the estimation algorithm. The optimal $Q(\boldsymbol{\mu}, \boldsymbol{\alpha}, \beta, \mathbf{x}_d)$ function can be obtained by minimizing the *free energy* function defined as [16]:

$$\mathbb{F} = \int_{\boldsymbol{\mu}, \boldsymbol{\alpha}, \beta, \mathbf{x}_d} Q(\boldsymbol{\mu}, \boldsymbol{\alpha}, \beta, \mathbf{x}_d) \log \frac{Q(\boldsymbol{\mu}, \boldsymbol{\alpha}, \beta, \mathbf{x}_d)}{p(\boldsymbol{\mu}, \boldsymbol{\alpha}, \beta, \mathbf{x}_d, \mathbf{y})} d\boldsymbol{\mu} d\boldsymbol{\alpha} d\beta d\mathbf{x}_d. \tag{24}$$

Notice that, $p(\boldsymbol{\mu}, \boldsymbol{\alpha}, \beta, \mathbf{x}_d, \mathbf{y})$ is used instead of $p(\boldsymbol{\mu}, \boldsymbol{\alpha}, \beta, \mathbf{x}_d|\mathbf{y})$ because they are proportional and thus equivalent in free energy formulation. According to the mean-field approximation [17], $Q(\boldsymbol{\mu}, \boldsymbol{\alpha}, \beta, \mathbf{x}_d)$ can be factorized into a product form, i.e., $Q(\boldsymbol{\mu}, \boldsymbol{\alpha}, \beta, \mathbf{x}_d) = Q(\boldsymbol{\mu})Q(\boldsymbol{\alpha})Q(\beta)Q(\mathbf{x}_d)$. This is equivalent to assuming that $\boldsymbol{\mu}, \boldsymbol{\alpha}, \beta$ and $\mathbf{x}_d$ are independent conditioned on $\mathbf{y}$ and will greatly simplify the iterative algorithm. With the mean-field approximation, the variational free energy in (24) is given by

$$\begin{aligned}
\mathbb{F} &= \int_{\boldsymbol{\mu}, \boldsymbol{\alpha}, \beta, \mathbf{x}_d} Q(\boldsymbol{\mu}, \boldsymbol{\alpha}, \beta, \mathbf{x}_d) \log \frac{Q(\boldsymbol{\mu}, \boldsymbol{\alpha}, \beta, \mathbf{x}_d)}{p(\mathbf{y}|\boldsymbol{\mu}, \beta, \mathbf{x}_d)p(\boldsymbol{\mu}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})p(\beta)p(\mathbf{x}_d)} d\boldsymbol{\mu} d\boldsymbol{\alpha} d\beta d\mathbf{x}_d \\
&= \int_{\boldsymbol{\mu}} Q(\boldsymbol{\mu}) \log Q(\boldsymbol{\mu}) d\boldsymbol{\mu} + \int_{\boldsymbol{\alpha}} Q(\boldsymbol{\alpha}) \log Q(\boldsymbol{\alpha}) d\boldsymbol{\alpha} + \int_{\beta} Q(\beta) \log Q(\beta) d\beta \\
&\quad + \int_{\mathbf{x}_d} Q(\mathbf{x}_d) \log Q(\mathbf{x}_d) d\mathbf{x}_d - \int_{\boldsymbol{\mu}, \boldsymbol{\alpha}} Q(\boldsymbol{\mu})Q(\boldsymbol{\alpha}) \log p(\boldsymbol{\mu}|\boldsymbol{\alpha}) d\boldsymbol{\mu} d\boldsymbol{\alpha} - \int_{\boldsymbol{\alpha}} Q(\boldsymbol{\alpha}) \log p(\boldsymbol{\alpha}) d\boldsymbol{\alpha} \\
&\quad - \int_{\beta} Q(\beta) \log p(\beta) d\beta - \int_{\mathbf{x}_d} Q(\mathbf{x}_d) \log p(\mathbf{x}_d) d\mathbf{x}_d \\
&\quad - \int_{\boldsymbol{\mu}, \beta, \mathbf{x}_d} Q(\boldsymbol{\mu})Q(\beta)Q(\mathbf{x}_d) \log p(\mathbf{y}|\boldsymbol{\mu}, \beta, \mathbf{x}_d) d\boldsymbol{\mu} d\beta d\mathbf{x}_d.
\end{aligned} \tag{25}$$

In order to calculate the free energy function given in (25), the likelihood function $p(\mathbf{y}|\boldsymbol{\mu}, \beta, \mathbf{x}_d)$ and the form of Q functions are needed, in addition to (20), (21), (22) and (23). Since the noise is assumed to be AWGN, the likelihood function is given by

$$p(\mathbf{y}|\boldsymbol{\mu}, \beta, \mathbf{x}_d) = \left(\frac{\beta}{\pi}\right)^N \exp\{-\beta(\mathbf{y} - \mathbf{G}[\mathbf{x}]\boldsymbol{\mu})^H(\mathbf{y} - \mathbf{G}[\mathbf{x}]\boldsymbol{\mu})\}. \tag{26}$$

For $Q(\mathbf{h})$, $Q(\boldsymbol{\alpha})$, $Q(\beta)$ and $Q(\mathbf{x}_d)$, they represent the approximate posterior distributions for the respective parameters, and should be chosen in a way that facilitate the manipulation. In particular, in order to maintain the sparsity enhancing property in the approximate posterior distribution of the channel BEM coefficients, we choose [10]

$$\begin{aligned} Q(\boldsymbol{\mu}) &= \frac{1}{\pi^M \det(\tilde{\boldsymbol{\Sigma}}_\mu)} \exp\{-(\boldsymbol{\mu} - \tilde{\mathbf{m}}_\mu)^H \tilde{\boldsymbol{\Sigma}}_\mu^{-1} (\boldsymbol{\mu} - \tilde{\mathbf{m}}_\mu)\} & (27) \\ Q(\alpha_j) &= Ga(\alpha_j | \tilde{a}_j, \tilde{b}_j) = \tilde{b}_j^{\tilde{a}_j} \alpha_j^{\tilde{a}_j - 1} \exp(-\tilde{b}_j \alpha_j)/\Gamma(\tilde{a}_j) & (28) \end{aligned}$$

with $\tilde{\mathbf{m}}_\mu$, $\tilde{\boldsymbol{\Sigma}}_\mu$ $\tilde{a}_j$ and $\tilde{b}_j$ being unknown parameters. Furthermore, for composite noise power, we set $Q(\beta)$ as

$$Q(\beta) = Ga(\beta | \tilde{c}, \tilde{d}) = \tilde{d}^{\tilde{c}} \beta^{\tilde{c} - 1} \exp(-\tilde{d}\beta)/\Gamma(\tilde{c}) \tag{29}$$

with $\tilde{c}$ and $\tilde{d}$ being unknown parameters. And for $\mathbf{x}_d$, in view of its discrete property, a close approximation is given as [18]

$$Q(\mathbf{x}_d) = \delta(\mathbf{x}_d - \tilde{\mathbf{x}}_d), \tag{30}$$

with $\tilde{\mathbf{x}}_d$ being a parameter of $Q(\mathbf{x}_d)$.

With all the distribution functions given above, the nine terms in (25) can be computed respectively. The detailed calculations are shown in the Appendix A. With the obtained results (49), (50), (51), (52), (54), (55), (56), (57), (58), and after eliminating some constant terms, the closed-form expression of the free energy function can be written as

$$\begin{aligned} & \mathbb{F}(\tilde{\mathbf{m}}_\mu, \tilde{\boldsymbol{\Sigma}}_\mu, \tilde{a}_j, \tilde{b}_j, \tilde{c}, \tilde{d}, \tilde{\mathbf{x}}_d) \\ = \; & -\log\det(\tilde{\boldsymbol{\Sigma}}_\mu) + \mathrm{Tr}\Big\{ \mathrm{diag}\big\{ [\frac{\tilde{a}_1}{\tilde{b}_1}, \ldots, \frac{\tilde{a}_M}{\tilde{b}_M}] \big\} (\tilde{\mathbf{m}}_\mu \tilde{\mathbf{m}}_\mu^H + \tilde{\boldsymbol{\Sigma}}_\mu) \Big\} \\ & + \sum_{j=1}^{M} \Big[\tilde{a}_j \log \tilde{b}_j + (\tilde{a}_j - 1)[\Psi(\tilde{a}_j) - \log \tilde{b}_j] - \tilde{a}_j - \log\Gamma(\tilde{a}_j) \Big] \\ & - \sum_{j=1}^{M} \Big[a_j \log b_j + (a_j - 1)[\Psi(\tilde{a}_j) - \log \tilde{b}_j] - b_j \tilde{a}_j / \tilde{b}_j - \log\Gamma(a_j) \Big] \\ & - \sum_{j=1}^{M} \Big[\Psi(\tilde{a}_j) - \log \tilde{b}_j \Big] + \tilde{c} \log \tilde{d} + (\tilde{c} - 1)[\Psi(\tilde{c}) - \log \tilde{d}] - \tilde{c} \\ & - \log\Gamma(\tilde{c}) - (c - 1)[\Psi(\tilde{c}) - \log \tilde{d}] + d\tilde{c}/\tilde{d} - N[\Psi(\tilde{c}) - \log \tilde{d}] \\ & + \frac{\tilde{c}}{\tilde{d}} \Big[\mathrm{Tr}\Big\{ \mathbf{G}^H[\tilde{\mathbf{x}}]\mathbf{G}[\tilde{\mathbf{x}}](\tilde{\mathbf{m}}_\mu \tilde{\mathbf{m}}_\mu^H + \tilde{\boldsymbol{\Sigma}}_\mu) \Big\} + \mathbf{y}^H \mathbf{y} - 2\mathrm{Re}\left\{ \mathbf{y}^H \mathbf{G}[\tilde{\mathbf{x}}] \tilde{\mathbf{m}}_\mu \right\} \Big] \\ & + \sum_{n=1}^{N_d} \log \Big\{ \sum_{\bar{x}_d(n) \in \mathbb{C}_d} \delta(\tilde{x}_d(n) - \bar{x}_d(n)) \Big\}, \qquad (31) \end{aligned}$$

where $\Psi(a) = \frac{\partial}{\partial a} \log \Gamma(a)$ is the digamma function, and we let $\tilde{\mathbf{x}} = \mathbf{E}_p \mathbf{x}_p + \mathbf{E}_d \tilde{\mathbf{x}}_d$ to simplify the expression. Notice that, the free energy function only depends on $\tilde{\mathbf{m}}_\mu$, $\tilde{\mathbf{\Sigma}}_\mu$, $\tilde{a}_j$, $\tilde{b}_j$, $\tilde{c}$, $\tilde{d}$ and $\tilde{\mathbf{x}}_d$.

C. Iterative Minimization of Free Energy Function

After obtaining the closed-form free energy function (31), the next step is to minimize the free energy function in order to obtain the optimal $\tilde{\mathbf{m}}_\mu, \tilde{\mathbf{\Sigma}}_\mu, \tilde{a}_j, \tilde{b}_j, \tilde{c}, \tilde{d}$ and $\tilde{\mathbf{x}}_d$. As the function depends on a large number of parameters, it is difficult to obtain the optimal parameters analytically in one step. The commonly used solution is to update each one in turn. In the following, it is shown that the closed-form solutions of $\tilde{\mathbf{m}}_\mu$ and $\tilde{\mathbf{\Sigma}}_\mu$ can be derived if the other parameters are fixed. Since the $Q(\boldsymbol{\mu})$ is assumed to be in Gaussian form, the optimal BEM coefficient estimate is equal to the mean, i.e., $\tilde{\mathbf{m}}_\mu$. Similarly, it is shown below that $(\tilde{a}_j, \tilde{b}_j)$ and $(\tilde{c}, \tilde{d})$ can be updated in pairs. Furthermore, we can derive the optimal $\tilde{\mathbf{x}}_d$ with Viterbi algorithm when other parameters are fixed. Therefore, $\mathbb{F}(\tilde{\mathbf{m}}_\mu, \tilde{\mathbf{\Sigma}}_\mu, \tilde{a}_j, \tilde{b}_j, \tilde{c}, \tilde{d}, \tilde{\mathbf{x}}_d)$ is minimized iteratively, starting with a certain initial value, and is guaranteed to converge [19].

1) Minimization w.r.t. $\tilde{\mathbf{x}}_d$: Gathering the terms in (31) that involve $\tilde{\mathbf{x}}_d$, we have

$$\begin{aligned}\mathbb{F}_{\tilde{x}_d} &= \frac{\tilde{c}}{\tilde{d}} \Big[\mathrm{Tr}\Big\{ \mathbf{G}^H[\tilde{\mathbf{x}}] \mathbf{G}[\tilde{\mathbf{x}}] (\tilde{\mathbf{m}}_\mu \tilde{\mathbf{m}}_\mu^H + \tilde{\mathbf{\Sigma}}_\mu) \Big\} - 2\mathrm{Re}\{ \mathbf{y}^H \mathbf{G}[\tilde{\mathbf{x}}] \tilde{\mathbf{m}}_\mu \} \Big] \\ &\quad + \sum_{n=1}^{N_d} \log \Big\{ \sum_{\bar{x}_d(n) \in \mathbb{C}_d} \delta(\tilde{x}_d(n) - \bar{x}_d(n)) \Big\}. \end{aligned} \tag{32}$$

Instead of treating $\tilde{\mathbf{x}}_d$ as continuous and taking derivatives of (32), which does not guarantee the resulted $\tilde{\mathbf{x}}_d$ will fall on the pre-defined constellation points, optimal $\tilde{\mathbf{x}}_d$ is searched on the constellation to minimize (32) as follows.

First, it should be noticed that if we only search on the constellation points, then $\tilde{x}_d(n) \in \mathbb{C}_d$ and $\sum_{\bar{x}_d(n) \in \mathbb{C}_d} \delta(\tilde{x}_d(n) - \bar{x}_d(n)) = 1$ for all n. Thus

$$\sum_{n=1}^{N_d} \log \Big\{ \sum_{\bar{x}_d(n) \in \mathbb{C}_d} \delta(\tilde{x}_d(n) - \bar{x}_d(n)) \Big\} = \sum_{n=1}^{N_d} \log\{1\} = 0. \tag{33}$$

Moreover, $\tilde{c}, \tilde{d} > 0$ from the property of Gamma distribution, then the factor $\tilde{c}/\tilde{d}$ can be excluded in the searching metric, and we have

$$\mathbb{F}_{\tilde{x}_d} = \mathrm{Tr}\Big\{ \mathbf{G}^H[\tilde{\mathbf{x}}] \mathbf{G}[\tilde{\mathbf{x}}] (\tilde{\mathbf{m}}_\mu \tilde{\mathbf{m}}_\mu^H + \tilde{\mathbf{\Sigma}}_\mu) \Big\} - 2\mathrm{Re}\{ \mathbf{y}^H \mathbf{G}[\tilde{\mathbf{x}}] \tilde{\mathbf{m}}_\mu \}. \tag{34}$$

It should be noticed that, the objective function given in (34) depends on $\tilde{\mathbf{x}}_d$ in a highly nonlinear way, making it difficult to find a solution for the optimal $\mathbf{x}_d$. In order to proceed, we perform the eigen-decomposition $\tilde{\mathbf{\Sigma}}_\mu = \sum_{j=1}^{M} \lambda_j \boldsymbol{\xi}_j \boldsymbol{\xi}_j^H$, and we have

$$\mathrm{Tr}\left\{\mathbf{G}^H[\tilde{\mathbf{x}}]\mathbf{G}[\tilde{\mathbf{x}}]\tilde{\mathbf{\Sigma}}_\mu\right\} = \sum_{j=1}^{M} \lambda_j \boldsymbol{\xi}_j^H \mathbf{G}^H[\tilde{\mathbf{x}}]\mathbf{G}[\tilde{\mathbf{x}}]\boldsymbol{\xi}_j. \tag{35}$$

Putting (35) into (34), we obtain

$$\mathbb{F}_{\tilde{x}_d} = \tilde{\mathbf{m}}_\mu^H \mathbf{G}^H[\tilde{\mathbf{x}}]\mathbf{G}[\tilde{\mathbf{x}}]\tilde{\mathbf{m}}_\mu + \sum_{j=1}^{M} \lambda_j \boldsymbol{\xi}_j^H \mathbf{G}^H[\tilde{\mathbf{x}}]\mathbf{G}[\tilde{\mathbf{x}}]\boldsymbol{\xi}_j - 2\mathrm{Re}\{\mathbf{y}^H \mathbf{G}_{\mathbf{h}}[\tilde{\mathbf{x}}]\tilde{\mathbf{m}}_\mu\}. \tag{36}$$

Due to the equality $\mathbf{G}[\mathbf{x}]\boldsymbol{\mu} = \mathbf{D}[\boldsymbol{\mu}]\mathbf{x}$ derived from (18) and (19), (36) can be written as

$$\mathbb{F}_{\tilde{x}_d} = \tilde{\mathbf{x}}^H \mathbf{D}^H[\tilde{\mathbf{m}}_\mu]\mathbf{D}[\tilde{\mathbf{m}}_\mu]\tilde{\mathbf{x}} + \sum_{j=1}^{M} \lambda_j \tilde{\mathbf{x}}^H \mathbf{D}^H[\boldsymbol{\xi}_j]\mathbf{D}[\boldsymbol{\xi}_j]\tilde{\mathbf{x}} - 2\mathrm{Re}\{\mathbf{y}^H \mathbf{D}[\tilde{\mathbf{m}}_\mu]\tilde{\mathbf{x}}\}. \tag{37}$$

Then we adopt the Viterbi algorithm [20] to minimize (37). The frequency domain signal $\tilde{\mathbf{x}}$ is treated as a sequence of data and the correlation between data is determined by the ICI. Strictly speaking, for DSC, all the elements of $\mathbf{D}[\tilde{\mathbf{m}}_\mu]$ and $\mathbf{D}[\boldsymbol{\xi}_j]$, $j = 1, \ldots, M$ are nonzero. But the entries close to the main diagonal are more prominent compared to those further away from the diagonal. This means that, the matrices $\mathbf{D}[\tilde{\mathbf{m}}_\mu]$ and $\mathbf{D}[\boldsymbol{\xi}_j]$ can be approximated as a banded matrix of bandwidth $B = 2\kappa + 1$, as shown in Figure 2. This approximated banded structure of DSC matrix considers the most significant ICI from the left and right κ closest symbols, and has been widely used in the literature [6], [20]. As shown in Figure 2, the upper right and lower left nonzero entries of the matrix cause cyclic interference between the first and last subcarriers, making the problem different from the uni-directional convolutional code decoding, where Viterbi algorithm is commonly used. In order to avoid the complication, we set the first and last κ symbols to be zero.

With the banded structure of $\mathbf{D}[\tilde{\mathbf{m}}_\mu]$ and $\mathbf{D}[\boldsymbol{\xi}_j]$, the branch metric becomes

$$\begin{aligned} \mathcal{V}_B([\tilde{\mathbf{x}}]_{n-2\kappa:n}; y(n-\kappa)) \;=\; & ([\mathbf{D}[\tilde{\mathbf{m}}_\mu]]_{n,n-\kappa:n-2\kappa}[\tilde{\mathbf{x}}]_{n-2\kappa:n})^H([\mathbf{D}[\tilde{\mathbf{m}}_\mu]]_{n,n-\kappa:n-2\kappa}[\tilde{\mathbf{x}}]_{n-2\kappa:n}) \\ & + \sum_{j=1}^{M} \lambda_j([\mathbf{D}[\boldsymbol{\xi}_j]]_{n,n-\kappa:n-2\kappa}[\tilde{\mathbf{x}}]_{n-2\kappa:n})^H([\mathbf{D}[\boldsymbol{\xi}_j]]_{n,n-\kappa:n-2\kappa}[\tilde{\mathbf{x}}]_{n-2\kappa:n}) \\ & -2\mathrm{Re}\{y^*(n-\kappa)[\mathbf{D}[\tilde{\mathbf{m}}_\mu]]_{n,n-\kappa:n-2\kappa}[\tilde{\mathbf{x}}]_{n-2\kappa:n}\}. \end{aligned} \tag{38}$$

Since known pilot symbols are inserted between unknown data, slight modification to standard Viterbi algorithm is needed. For $n \in \mathfrak{I}_p$, the true pilot value will be used to calculate the branch value, and the number of remaining paths will reduce by half because of merging. And for $n \in \mathfrak{I}_d$, the algorithm will perform as normal Viterbi algorithm.

2) Minimization w.r.t. $\tilde{\mathbf{m}}_\mu, \tilde{\mathbf{\Sigma}}_\mu, \tilde{a}_j, \tilde{b}_j, \tilde{c}$ *and* $\tilde{d}$*:* The optimal values of other unknown parameters are obtained by setting the first order derivative of (31) with respect to the corresponding parameter to zero. As shown in Appendix B, we obtain the following set of solutions:

$$\begin{aligned}
\tilde{\mathbf{\Sigma}}_\mu &= \left(\mathrm{diag}\left\{\left[\frac{\tilde{a}_1}{\tilde{b}_1}, \ldots, \frac{\tilde{a}_M}{\tilde{b}_M}\right]\right\} + \frac{\tilde{c}}{\tilde{d}}\mathbf{G}^H[\tilde{\mathbf{x}}]\mathbf{G}[\tilde{\mathbf{x}}]\right)^{-1} && (39)\\
\tilde{\mathbf{m}}_\mu &= \frac{\tilde{c}}{\tilde{d}}\tilde{\mathbf{\Sigma}}_\mu\mathbf{G}^H[\tilde{\mathbf{x}}]\mathbf{y} && (40)\\
\tilde{a}_j &= a_j + 1 && (41)\\
\tilde{b}_j &= b_j + |[\tilde{\mathbf{m}}_\mu]_j|^2 + [\tilde{\mathbf{\Sigma}}_\mu]_{j,j} && (42)\\
\tilde{c} &= c + N && (43)\\
\tilde{d} &= d + \mathbf{y}^H\mathbf{y} - 2\mathrm{Re}\left\{\mathbf{y}^H\mathbf{G}[\tilde{\mathbf{x}}]\tilde{\mathbf{m}}_\mu\right\} + \mathrm{Tr}\left\{\mathbf{G}^H[\tilde{\mathbf{x}}]\mathbf{G}[\tilde{\mathbf{x}}]\left(\tilde{\mathbf{m}}_\mu\tilde{\mathbf{m}}_\mu^H + \tilde{\mathbf{\Sigma}}_\mu\right)\right\}. && (44)
\end{aligned}$$

It is worth noting that, along with each update, when $|[\tilde{\mathbf{m}}_\mu]_j|^2 + [\tilde{\mathbf{\Sigma}}_\mu]_{j,j}$ gets close to zero, meaning both mean and variance of the corresponding μ_j are close to zero, then μ_j can be treated as null entry and pruned from further iteration. In practice, a threshold with the order of 10^{-10} is used to compare with $|[\tilde{\mathbf{m}}_\mu]_j|^2 + [\tilde{\mathbf{\Sigma}}_\mu]_{j,j}$ to determine which μ_j is being pruned [21].

D. Summary of the Iterative Algorithm

We summarize the parameter updating procedure as follows

Initialization: Choose initial values $\{\tilde{a}_1^0, \ldots, \tilde{a}_M^0\}$, $\{\tilde{b}_1^0, \ldots, \tilde{b}_M^0\}$, $\tilde{c}^0, \tilde{d}^0$ and $\tilde{\mathbf{x}}_d^0$.

Iterations: For the i^{th} iteration

Updating the parameters of GCE-BEM coefficients

$$\begin{aligned}
\tilde{\mathbf{\Sigma}}_\mu^i &= \left(\mathrm{diag}\{[\frac{\tilde{a}_1^{i-1}}{\tilde{b}_1^{i-1}}, \ldots, \frac{\tilde{a}_M^{i-1}}{\tilde{b}_M^{i-1}}]\} + \frac{\tilde{c}^{i-1}}{\tilde{d}^{i-1}}\mathbf{G}^H[\tilde{\mathbf{x}}^{i-1}]\mathbf{G}[\tilde{\mathbf{x}}^{i-1}]\right)^{-1}\\
\tilde{\mathbf{m}}_\mu^i &= \frac{\tilde{c}^{i-1}}{\tilde{d}^{i-1}}\tilde{\mathbf{\Sigma}}_\mu^i\mathbf{G}^H[\tilde{\mathbf{x}}^{i-1}]\mathbf{y}
\end{aligned}$$

Updating the hyperparameters of GCE-BEM coefficients

$$\begin{aligned}
\tilde{a}_j^i &= a_j + 1\\
\tilde{b}_j^i &= b_j + |[\tilde{\mathbf{m}}_\mu]_j|^2 + [\tilde{\mathbf{\Sigma}}_\mu]_{j,j}
\end{aligned}$$

Updating the estimate of data

$$\min_{\tilde{\mathbf{x}}_d} \mathbb{F}_{\tilde{x}_d} = -\,2\mathrm{Re}\left\{\mathbf{y}^H\mathbf{D}[\tilde{\mathbf{m}}_\mu]\tilde{\mathbf{x}}\right\} + \tilde{\mathbf{x}}^H\mathbf{D}^H[\tilde{\mathbf{m}}_\mu]\mathbf{D}[\tilde{\mathbf{m}}_\mu]\tilde{\mathbf{x}} + \sum_{j=1}^{M}\lambda_j\tilde{\mathbf{x}}^H\mathbf{D}^H[\boldsymbol{\xi}_j]\mathbf{D}[\boldsymbol{\xi}_j]\tilde{\mathbf{x}},$$

where $\tilde{\boldsymbol{\Sigma}}_\mu^i = \sum_{j=1}^{M^{i-1}} \lambda_j^i \boldsymbol{\xi}_j^i(\boldsymbol{\xi}_j^i)^H$, using Viterbi algorithm.

Updating the hyperparameters of noise

$$\begin{aligned}\tilde{c}^i &= c + N\\ \tilde{d}^i &= d + \mathbf{y}^H\mathbf{y} - 2\mathrm{Re}\left\{\mathbf{y}^H\mathbf{G}[\tilde{\mathbf{x}}^i]\tilde{\mathbf{m}}_\mu^i\right\} + \mathrm{Tr}\left\{\mathbf{G}^H[\tilde{\mathbf{x}}^i]\mathbf{G}[\tilde{\mathbf{x}}^i]\left(\tilde{\mathbf{m}}_\mu^i(\tilde{\mathbf{m}}_\mu^i)^H + \tilde{\boldsymbol{\Sigma}}_\mu^i\right)\right\}\end{aligned}$$

Pruning

If $|[\tilde{\mathbf{m}}_\mu]_j|^2 + [\tilde{\boldsymbol{\Sigma}}_\mu]_{j,j} < 10^{-10} \Rightarrow \mu_j = 0$, remove $\tilde{a}_j$ from $\{\tilde{a}_1^i, \dots, \tilde{a}_M^i\}$ and $\tilde{b}_j$ from $\{\tilde{b}_1^i, \dots, \tilde{b}_M^i\}$ and update the expression of $\mathbf{G}[\tilde{\mathbf{x}}^i]$ by removing the j^{th} column.

End

It is worth noting that the values of $\{\tilde{a}_1, \dots, \tilde{a}_M\}$ and $\tilde{c}$ remain the same for each iteration, thus they should only be updated in the first iteration in practice.

V. Initialization of the Iterative Algorithm

In order to start the iteration, one of the quantities we need is the initial data estimate $\tilde{\mathbf{x}}_d^0$. If an initial channel estimate can be obtained from pilots, then from (19), the initial data estimate can be obtained. In Section III, GCE-BEM with large oversampling factor is used to represent the channel. Though this provides flexibility for selecting important bases in the iterative channel estimation and data detection, the large number of unknown coefficients corresponding to GCE-BEM also brings challenges in the initial value estimation, which is important to any iterative algorithm. Unfortunately, traditional compressed sensing methods like BP and orthogonal matching pursuit (OMP), whose performance highly depends on the accurate knowledge of noise variance, is not applicable in our case, since the variance of the composite noise is unknown.

On the other hand, notice that the proposed iterative algorithm does not rely directly on the estimate of $\boldsymbol{\mu}$ to start the iteration. The channel is needed only indirectly through initial data estimate. Thus, during initial data detection, we choose to expand the channel with CE-BEM, which corresponds to choosing the oversampling factor of GCE-BEM as $V = 1$. Then least squares (LS) algorithm is used to obtain the initial channel estimation. Though CE-BEM is widely reported as relatively inaccurate among other BEMs, it represents the time-varying channel using a very small number of orthogonal bases. This is highly beneficial for a rough estimate at the initial stage where there is little knowledge of the channel and source signal at the receiver.

According to (2), $\mathbf{x} = \mathbf{E}_d\mathbf{x}_d + \mathbf{E}_p\mathbf{x}_p$. Together with the fact that $\mathbf{G}[\mathbf{E}_p\mathbf{x}_p + \mathbf{E}_d\mathbf{x}_d] = \mathbf{G}[\mathbf{E}_p\mathbf{x}_p] + \mathbf{G}[\mathbf{E}_d\mathbf{x}_d]$, (18) can be written as $\mathbf{y} = \mathbf{G}[\mathbf{E}_p\mathbf{x}_p]\boldsymbol{\mu} + \mathbf{G}[\mathbf{E}_d\mathbf{x}_d]\boldsymbol{\mu} + \mathbf{v}$. Collecting the output samples corresponding to pilot positions $\mathfrak{I}_p$, the equation for initial channel estimation can be written as

$$\mathbf{y}_p = \mathbf{G}_p[\mathbf{E}_p\mathbf{x}_p]\boldsymbol{\mu} + \mathbf{G}_p[\mathbf{E}_d\mathbf{x}_d]\boldsymbol{\mu} + \mathbf{v}_p, \tag{45}$$

where $\mathbf{G}_p[.]$ is constructed from the rows of $\mathbf{G}[.]$ corresponding to $\mathfrak{I}_p$. The initial channel estimation can be obtained by treating the second term of (45) as noise and performing LS algorithm:

$$\hat{\boldsymbol{\mu}} = (\mathbf{G}_p^H[\mathbf{E}_p\mathbf{x}_p]\mathbf{G}_p[\mathbf{E}_p\mathbf{x}_p])^{-1}\mathbf{G}_p^H[\mathbf{E}_p\mathbf{x}_p]\mathbf{y}_p. \tag{46}$$

With the estimated channel $\hat{\boldsymbol{\mu}}$, we rewrite (19) as $\mathbf{y} = \mathbf{D}[\hat{\boldsymbol{\mu}}]\mathbf{E}_d\mathbf{x}_d + \mathbf{D}[\hat{\boldsymbol{\mu}}]\mathbf{E}_p\mathbf{x}_p + \mathbf{v}$. Applying LS estimation again, we have

$$\hat{\mathbf{x}}_d^0 = (\mathbf{E}_d^H\mathbf{D}^H[\hat{\boldsymbol{\mu}}]\mathbf{D}[\hat{\boldsymbol{\mu}}]\mathbf{E}_d)^{-1}\mathbf{E}_d^H\mathbf{D}^H[\hat{\boldsymbol{\mu}}](\mathbf{y} - \mathbf{D}[\hat{\boldsymbol{\mu}}]\mathbf{E}_p\mathbf{x}_p). \tag{47}$$

The obtained $\hat{\mathbf{x}}_d^0$ may not reside on the constellation map, thus quantization is performed on $\hat{\mathbf{x}}_d^0$ and the initial data detection is given as $\tilde{\mathbf{x}}_d^0 = \mathrm{Qant}[\hat{\mathbf{x}}_d^0]$. Notice that in DSC, the ICI is not negligible, and $\mathbf{G}_p[\mathbf{E}_d\mathbf{x}_d]\boldsymbol{\mu} \neq \mathbf{0}$, which decreases the accuracy of estimation in (46), and in turns affects the accuracy of initial data detection. This is the reason why an iterative algorithm is necessary.

For other initial values $\{\tilde{a}_1^0, \ldots, \tilde{a}_M^0\}$, $\{\tilde{b}_1^0, \ldots, \tilde{b}_M^0\}$, $\tilde{c}^0, \tilde{d}^0$ in the iterative algorithm, it is should be noticed that only the ratios $\tilde{a}_j/\tilde{b}_j$ and $\tilde{c}/\tilde{d}$ are required, thus we only need to specify the initial values of the ratios to start the iteration. From (28) and the property of Gamma distribution, $\tilde{a}_j/\tilde{b}_j$ represents the mean value of α_j, which is the inverse variance of channel GCE-BEM coefficients. Since we have no information about their relative values, we can set them to be equal. That is, let $\tilde{a}_j^0/\tilde{b}_j^0 = 1/M$ for all j. Furthermore, from (29) and the property of Gamma distribution, $\tilde{c}/\tilde{d} = \mathbb{E}\{\beta\} = \mathbb{E}\{1/\varpi_v^2\}$. Therefore the initial value can be set as $\tilde{c}^0/\tilde{d}^0 = 1/\hat{\varpi}_v^2$, where $\hat{\varpi}_v^2$ is an estimate of noise power

$$\hat{\varpi}_v^2 = \mathbf{y} - \mathbf{G}[\mathbf{E}_p\mathbf{x}_p + \mathbf{E}_d\tilde{\mathbf{x}}_d^0]\hat{\boldsymbol{\mu}}. \tag{48}$$

VI. Simulation Results and Discussions

In this section, simulation results of dual-hop and three-hop cooperative OFDM systems are provided. In both systems, each OFDM symbol has 128 subcarriers and the length of CP is 8. Carrier frequency is $f_c = 2$GHz and the sample interval is $T_s = 2\mu s$. The channel $\bar{h}_{k,\rho}(n,l)$ is generated according to zero-mean complex Gaussian distribution with autocorrelation of the l^{th} tap given by $\mathbb{E}\{\bar{h}_{k,\rho}(m,l)\bar{h}_{k,\rho}(n,l)\} = \sigma^2_{k,\rho}(l)J_0(2\pi f_{k,\rho}(l)(m-n)T_s)$ [22], where $J_0(\cdot)$ represents the zero-order Bessel function of the first kind, and $\sigma^2_{k,\rho}(l)$ is the power of the l^{th} tap. Fourteen pilot clusters are used. The clusters are equal-spaced and interleaved with data subcarriers. In each cluster, one nonzero pilot is guarded by one zero pilot on each side. The nonzero pilots are generated as zero-mean complex Gaussian random variables with power three times that of data symbols. And the data is modulated with QPSK of unit power. The normalized channel mean-square error (MSE) and data detection bit error rate (BER) are plotted to demonstrate the performance. The MSE of the channel estimate at the i^{th} iteration is defined as $\mathrm{MSE}^i = ||\hat{\mathbf{H}}^i - \mathbf{H}||^2/||\mathbf{H}||^2$, where $\hat{\mathbf{H}}^i$ is the channel matrix recovered from the GCE-BEM estimate at the i^{th} iteration. The noise power at the relays and destination are set to be the same $\varpi^2_d = \varpi^2_{k,\rho}$, for all k and ρ. The signal-to-noise ratio (SNR) in the following figures is defined as $\mathrm{SNR} = \sigma^2_s/\varpi^2_d$ [9]. The oversampling factor is chosen as $V = 20$ for GCE-BEM. Each point is obtained by averaging the results over 1,000 runs.

For the dual-hop system, two relaying paths (K = 2) are considered. For both relaying paths ($k = 1, 2$), the maximal normalized Doppler shifts[1] are set as 0.05 for the first hop and 0.15 for the second hop. In the simulation, for each specific channel, one randomly chosen tap has Doppler shift equals the maximum Doppler shift specified above. And for other taps, their Doppler shifts are uniformly drawn within the range from 0 to the maximum Doppler shift. Both source-relay channels have 2 taps. One of the source-relay channels has tap positions uniformly drawn from $\{0, 1, 2\}$ while the other has tap positions uniformly drawn from $\{0, 1, 2, 3, 4\}$. And both relay-destination channels have 2 taps, with the tap positions uniformly drawn from $\{0, 1, 2, 3\}$ for one channel and from $\{0, 1, 2\}$ for the other. All the channels follow exponential power delay profiles normalized to unit power. For the Viterbi equalizer, $\kappa = 3$ is chosen.

Figure 3 and Figure 4 present the convergence performance of the proposed iterative algorithm

[1] Normalized Doppler shift is defined as Nf_dT_s with f_d being the Doppler frequency.

in terms of MSE and BER, respectively. SNR is set at 10dB, 20dB and 30dB. It can be seen that the MSEs and BERs improve significantly in the first few iterations and converge to stable values before 10 iterations.

Figure 5 and Figure 6 show the MSE and BER performance achieved by the proposed iterative algorithm versus SNRs. The results are taken after 10 iterations in order to guarantee convergence. In the figures, KLEM represents the performance of the EM algorithm with channel expanded on Karhuen-Loève (KL) bases. This algorithm is an extension of the KLEM algorithm for single-hop case [23], and the detail is not included in this paper due to space limitation. The KLEM algorithm requires full information on channel tap positions, Doppler frequencies and power profile of each channel, together with noise statistics, thus serves as a reference for optimal performance here. And CRLB curve represents the Cramé-Rao lower bound, which can be obtained from [23] by replacing the single-hop channel and noise power with the composite channel and composite noise power. Meanwhile, ideal case with full channel information at the receiver is also depicted as the performance bound in the BER figure. From Figure 5 and 6, it can be seen that, the proposed iterative algorithm successfully eliminates the interdependence between data detection and channel estimation, and exhibits significant performance improvement compared to the initial channel estimation MSE and data detection BER. Furthermore, though the proposed iterative algorithm does not have access to the relaying system structure (number of available links K) or any statistical information of the channel and noise, there are only minor performance gaps between the proposed method and the KLEM. Furthermore, the proposed algorithm and KLEM are very close to the ideal data detection in terms of BER performance.

For three-hop relaying system, the maximal normalized Doppler shifts for the first and third hop are set as 0.05 while that of the second hop is set as 0.15. Two relaying paths are considered (K= 2). For the first relaying path, the number of channel taps in the three hops are $\{2, 3, 2\}$, respectively; while that for the second relaying path is $\{3, 2, 2\}$, respectively. The channel taps in each hop are consecutive. For the Viterbi equalizer, $\kappa = 4$ is chosen.

Figure 7 and Figure 8 show the MSE and BER performance achieved by the proposed iterative algorithm versus SNRs, with results taken after 10 iterations[2]. In both figures, performance

[2] As the convergence performance of a three-hop system is similar to that of dual-hop system, the convergence figures are not shown here.

curves of KLEM, which demands detailed information of relaying system structure, channel and noise statistics, are depicted as a reference for optimal channel estimation and data detection. From Figure 7, it is seen that, the proposed iterative algorithm greatly improve the performance from the initial channel estimation, indicating the ability of the proposed algorithm to cancel interference between unknown data and pilots through iterations. Furthermore, after convergence, only a small performance gap exists between the proposed algorithm and KLEM, which touches the CRLB at high SNRs. This exhibits the strong ability of our proposed algorithm in learning the statistics of both channel and noise. From Figure 8, the BER performance of our proposed method is also shown to improve significantly compared to the initial data detection and is very close that of KLEM algorithm. From these figures, it can be concluded that, though the system model in a three-hop system is more complicated than that of dual-hop case, the proposed algorithm continues to present good performance in terms of both channel estimation MSE and data detection BER and demonstrate robustness in a variety of OFDM relaying systems.

VII. Conclusions

In this paper, channel estimation and data detection for multihop OFDM relaying system under high mobility has been investigated with focus on unknown channel orders and Doppler frequencies. By exploring the matrix structure of channels in different hops, we first simplified the multihop multilink channel matrix into a composite channel matrix. Then the composite channel was represented using GCE-BEM with a large oversampling factor so that sparsity on the delay-Doppler domain was introduced. Sparsity enhancing Gaussian priors with Gamma hyperpriors were adopted to enable the identification of nonzero entries. A pilot-aided iterative algorithm was developed under variational inference (VI) framework, using only limited number of pilot subcarriers in one OFDM symbol. The proposed algorithm iteratively estimates the channel, recovers the unknown data using Viterbi algorithm and learns the channel and noise statistical information. Simulation results showed that, even without any specific information on system structure, channel tap positions, channel lengths, Doppler shifts and noise powers, the proposed algorithms exhibited performance very close to that of an optimal channel estimation and data detection algorithm, which requires all of the above information.

Appendix A
Calculation of Free Energy Function

Taking logarithm on (27) and substituting the result to the ***first term*** of (25), we obtain

$$\begin{aligned}
\int_{\boldsymbol{\mu}} Q(\boldsymbol{\mu}) \log Q(\boldsymbol{\mu}) d\boldsymbol{\mu} &= -M \log \pi - \log\det(\tilde{\boldsymbol{\Sigma}}_\mu) - \tilde{\mathbf{m}}_\mu^H \tilde{\boldsymbol{\Sigma}}_\mu^{-1} \tilde{\mathbf{m}}_\mu \\
&\quad + 2\mathrm{Re}\left\{\mathbb{E}\{\boldsymbol{\mu}^H\}\tilde{\boldsymbol{\Sigma}}_\mu^{-1}\tilde{\mathbf{m}}_\mu\right\} - \mathrm{Tr}\left\{\tilde{\boldsymbol{\Sigma}}_\mu^{-1}\mathbb{E}\{\boldsymbol{\mu}\boldsymbol{\mu}^H\}\right\} \\
&= -M \log \pi - \log\det(\tilde{\boldsymbol{\Sigma}}_\mu) - \tilde{\mathbf{m}}_\mu^H \tilde{\boldsymbol{\Sigma}}_\mu^{-1} \tilde{\mathbf{m}}_\mu \\
&\quad + 2\tilde{\mathbf{m}}_\mu^H \tilde{\boldsymbol{\Sigma}}_\mu^{-1} \tilde{\mathbf{m}}_\mu - \mathrm{Tr}\left\{\tilde{\boldsymbol{\Sigma}}_\mu^{-1}(\tilde{\mathbf{m}}_\mu \tilde{\mathbf{m}}_\mu^H + \tilde{\boldsymbol{\Sigma}}_\mu)\right\} \\
&= -M \log \pi - \log\det(\tilde{\boldsymbol{\Sigma}}_\mu) - M. \qquad (49)
\end{aligned}$$

On the other hand, from the form of $Q(\alpha_j)$ given in (28), and notice that $Q(\boldsymbol{\alpha}) = \prod_{j=1}^M Q(\alpha_j)$, we can compute the ***second term*** of (25) as

$$\begin{aligned}
\int_{\boldsymbol{\alpha}} Q(\boldsymbol{\alpha}) \log Q(\boldsymbol{\alpha}) d\boldsymbol{\alpha} &= \sum_{j=1}^M \left[\tilde{a}_j \log \tilde{b}_j - \log\Gamma(\tilde{a}_j) + (\tilde{a}_j - 1)\mathbb{E}\{\log \alpha_j\} - \tilde{b}_j \mathbb{E}\{\alpha_j\}\right] \\
&= \sum_{j=1}^M \Big[\tilde{a}_j \log \tilde{b}_j - \log\Gamma(\tilde{a}_j) \\
&\quad + (\tilde{a}_j - 1)\left(\Psi(\tilde{a}_j) - \log \tilde{b}_j\right) - \tilde{b}_j\left(\tilde{a}_j / \tilde{b}_j\right)\Big] \\
&= \sum_{j=1}^M \left[\log \tilde{b}_j - \log\Gamma(\tilde{a}_j) + (\tilde{a}_j - 1)\Psi(\tilde{a}_j) - \tilde{a}_j\right], \qquad (50)
\end{aligned}$$

where the digamma function Ψ is defined by $\Psi(a) = \dfrac{\partial}{\partial a}\log\Gamma(a)$. Furthermore, since $Q(\beta)$ is in the same form as $Q(\alpha_j)$, following a similar derivation as above, it can be easily shown that the ***third term*** of (25) is

$$\int_\beta Q(\beta) \log Q(\beta) d\beta = \log \tilde{d} - \log\Gamma(\tilde{c}) + (\tilde{c} - 1)\Psi(\tilde{c}) - \tilde{c}. \qquad (51)$$

Based on the Dirac delta function in (30), the ***forth term*** of (25) is given by

$$\int_{\mathbf{x}_d} Q(\mathbf{x}_d) \log Q(\mathbf{x}_d) d\mathbf{x}_d = \log Q(\tilde{\mathbf{x}}_d) = \log \delta(\tilde{\mathbf{x}}_d - \tilde{\mathbf{x}}_d) = 0. \qquad (52)$$

Furthermore, from (20), we have

$$\begin{aligned}
\log p(\boldsymbol{\mu}|\boldsymbol{\alpha}) &= -M \log \pi - \log\det(\mathbf{A}^{-1}) - \boldsymbol{\mu}^H \mathbf{A} \boldsymbol{\mu} \\
&= -M \log \pi + \log\left(\prod_{j=1}^M \alpha_j\right) - \mathrm{Tr}\left\{\mathbf{A}\boldsymbol{\mu}\boldsymbol{\mu}^H\right\}. \qquad (53)
\end{aligned}$$

And the ***fifth term*** of (25) can be computed as

$$\begin{aligned}
&\int_{\boldsymbol{\mu},\boldsymbol{\alpha}} Q(\boldsymbol{\mu})Q(\boldsymbol{\alpha})\log p(\boldsymbol{\mu}|\boldsymbol{\alpha})d\boldsymbol{\mu}d\boldsymbol{\alpha} \\
= &\ -M\log\pi + \sum_{j=1}^{M}\mathbb{E}_\alpha\{\log\alpha_j\} - \mathrm{Tr}\{\mathbb{E}_\alpha\{\mathrm{diag}\{\boldsymbol{\alpha}\}\}\mathbb{E}_\mu\{\boldsymbol{\mu}\boldsymbol{\mu}^H\}\} \\
= &\ -M\log\pi + \sum_{j=1}^{M}\Big(\Psi(\tilde{a}_j) - \log\tilde{b}_j\Big) - \mathrm{Tr}\Bigg\{\mathrm{diag}\left\{\left[\frac{\tilde{a}_1}{\tilde{b}_1},\ldots,\frac{\tilde{a}_M}{\tilde{b}_M}\right]\right\}\Big(\tilde{\mathbf{m}}_\mu\tilde{\mathbf{m}}_\mu^H + \tilde{\boldsymbol{\Sigma}}_\mu\Big)\Bigg\}
\end{aligned}\tag{54}$$

From (21), we can compute the logarithm of $p(\boldsymbol{\alpha}) = \prod_{j=1}^{M} p(\alpha_j)$, and the ***sixth term*** of (25) can be written as

$$\begin{aligned}
\int_{\boldsymbol{\alpha}} Q(\boldsymbol{\alpha})\log p(\boldsymbol{\alpha})d\boldsymbol{\alpha} &= \sum_{j=1}^{M}\big[a_j\log b_j - \log\Gamma(a_j) + (a_j-1)\mathbb{E}_\alpha\{\log\alpha_j\} - b_j\mathbb{E}\{\alpha_j\}\big] \\
&= \sum_{j=1}^{M}\big[a_j\log b_j - \log\Gamma(a_j) + (a_j-1)(\Psi(\tilde{a}_j) - \log\tilde{b}_j) - b_j\frac{\tilde{a}_j}{\tilde{b}_j}\big].
\end{aligned}\tag{55}$$

Similarly, as $Q(\beta)$ and $p(\beta)$ are in the same form as $Q(\alpha_j)$ and $p(\alpha_j)$, we can easily show that the ***seventh term*** of (25) is

$$\int_\beta Q(\beta)\log p(\beta)d\beta = c\log d - \log\Gamma(c) + (c-1)\Big(\Psi(\tilde{c}) - \log\tilde{d}\Big) - d\frac{\tilde{c}}{\tilde{d}}. \tag{56}$$

From (22), we can obtain the logarithm of $p(\mathbf{x}_d)$. Together with (30), the ***eighth term*** of (25) can be derived as

$$\begin{aligned}
\int_{\mathbf{x}_d} Q(\mathbf{x}_d)\log p(\mathbf{x}_d)d\mathbf{x}_d &= \int_{\mathbf{x}_d}\delta(\mathbf{x}_d - \tilde{\mathbf{x}}_d)\sum_{n=1}^{N_d}\log\big\{\sum_{\bar{x}_d(n)\in\mathbb{C}_d}\delta(x_d(n) - \bar{x}_d(n))\big\}d\mathbf{x}_d - \log\{\mathcal{M}_d^{N_d}\} \\
&= \sum_{n=1}^{N_d}\log\Big\{\sum_{\bar{x}_d(n)\in\mathbb{C}_d}\delta(\tilde{x}_d(n) - \bar{x}_d(n))\Big\} - N_d\log\{\mathcal{M}_d\}.
\end{aligned}\tag{57}$$

Finally, taking the logarithm of (26), we have the ***ninth term*** of (25) given by

$$\begin{aligned}
&\int_{\boldsymbol{\mu},\beta,\mathbf{x}_d} Q(\boldsymbol{\mu})Q(\beta)Q(\mathbf{x}_d)\log p(\mathbf{y}|\boldsymbol{\mu},\beta,\mathbf{x}_d)d\boldsymbol{\mu}d\beta d\mathbf{x}_d \\
= &\ -N\log\pi + N\mathbb{E}_\beta\{\log\beta\} - \mathbb{E}_\beta\{\beta\}\Big[\mathbf{y}^H\mathbf{y} - 2\mathrm{Re}\{\mathbf{y}^H\mathbf{G}[\mathbf{E}_p\mathbf{x}_p + \mathbf{E}_d\tilde{\mathbf{x}}_d]\mathbb{E}_\mu\{\boldsymbol{\mu}\}\} \\
&\ +\mathrm{Tr}\big\{\mathbf{G}^H[\mathbf{E}_p\mathbf{x}_p + \mathbf{E}_d\tilde{\mathbf{x}}_d]\mathbf{G}[\mathbf{E}_p\mathbf{x}_p + \mathbf{E}_d\tilde{\mathbf{x}}_d]\mathbb{E}_\mu\{\boldsymbol{\mu}\boldsymbol{\mu}^H\}\big\}\Big] \\
= &\ -N\log\pi + N\Big(\Psi(\tilde{c}) - \log(\tilde{d})\Big) - \left(\frac{\tilde{c}}{\tilde{d}}\right)\Big[\mathbf{y}^H\mathbf{y} - 2\mathrm{Re}\Big\{\mathbf{y}^H\mathbf{G}[\mathbf{E}_p\mathbf{x}_p + \mathbf{E}_d\tilde{\mathbf{x}}_d]\tilde{\mathbf{m}}_\mu\Big\} \\
&\ +\mathrm{Tr}\Big\{\mathbf{G}^H[\mathbf{E}_p + \mathbf{x}_p + \mathbf{E}_d\tilde{\mathbf{x}}_d]\mathbf{G}[\mathbf{E}_p\mathbf{x}_p + \mathbf{E}_d\tilde{\mathbf{x}}_d]\Big(\tilde{\mathbf{m}}_\mu\tilde{\mathbf{m}}_\mu^H + \tilde{\boldsymbol{\Sigma}}_\mu\Big)\Big\}\Big].
\end{aligned}\tag{58}$$

APPENDIX B

DERIVATION OF UPDATING FUNCTIONS

Updating $(\tilde{\mathbf{m}}_\mu, \tilde{\mathbf{\Sigma}}_\mu)$ given $\tilde{a}_j, \tilde{b}_j, \tilde{c}, \tilde{d}$ and $\tilde{\mathbf{x}}_d$

Focusing on the terms in $\mathbb{F}$ related to $\tilde{\mathbf{\Sigma}}_\mu$, we have

$$\begin{aligned}\frac{\partial \mathbb{F}}{\partial \tilde{\mathbf{\Sigma}}_\mu} &= \frac{\partial}{\partial \tilde{\mathbf{\Sigma}}_\mu}\left\{-\log\det(\tilde{\mathbf{\Sigma}}_\mu) + \mathrm{Tr}\{\mathrm{diag}\{[\frac{\tilde{a}_1}{\tilde{b}_1},\ldots,\frac{\tilde{a}_M}{\tilde{b}_M}]\}\tilde{\mathbf{\Sigma}}_\mu\} + \frac{\tilde{c}}{\tilde{d}}[\mathbf{G}^H[\tilde{\mathbf{x}}]\mathbf{G}[\tilde{\mathbf{x}}]\tilde{\mathbf{\Sigma}}_\mu]\right\} \\ &= -\tilde{\mathbf{\Sigma}}_\mu^{-1} + \mathrm{diag}\left\{\left[\frac{\tilde{a}_1}{\tilde{b}_1},\ldots,\frac{\tilde{a}_M}{\tilde{b}_M}\right]\right\} + \frac{\tilde{c}}{\tilde{d}}\mathbf{G}^H[\tilde{\mathbf{x}}]\mathbf{G}[\tilde{\mathbf{x}}]. \end{aligned} \tag{59}$$

Setting (59) to zero leads to (39). On the other hand,

$$\begin{aligned}\frac{\partial \mathbb{F}}{\partial \tilde{\mathbf{m}}_\mu} &= \frac{\partial}{\partial \tilde{\mathbf{m}}_\mu}\left\{\mathrm{Tr}\left\{\mathrm{diag}\left\{\left[\frac{\tilde{a}_1}{\tilde{b}_1},\ldots,\frac{\tilde{a}_M}{\tilde{b}_M}\right]\right\}\tilde{\mathbf{m}}_\mu\tilde{\mathbf{m}}_\mu^H\right\} + \frac{\tilde{c}}{\tilde{d}}\Big[-2\mathrm{Re}\left\{\mathbf{y}^H\mathbf{G}[\tilde{\mathbf{x}}]\tilde{\mathbf{m}}_\mu\right\}\right. \\ &\quad \left. + \mathbf{G}^H[\tilde{\mathbf{x}}]\mathbf{G}[\tilde{\mathbf{x}}]\tilde{\mathbf{m}}_\mu\tilde{\mathbf{m}}_\mu^H\Big]\right\} \\ &= \frac{\partial}{\partial \tilde{\mathbf{m}}_\mu}\left\{\tilde{\mathbf{m}}_\mu^H\left[\mathrm{diag}\left\{\left[\frac{\tilde{a}_1}{\tilde{b}_1},\ldots,\frac{\tilde{a}_M}{\tilde{b}_M}\right]\right\} + \frac{\tilde{c}}{\tilde{d}}\mathbf{G}^H\left[\tilde{\mathbf{x}}\right]\mathbf{G}[\tilde{\mathbf{x}}]\right]\tilde{\mathbf{m}}_\mu - \frac{\tilde{c}}{\tilde{d}}\tilde{\mathbf{m}}_\mu^H\mathbf{G}^H[\tilde{\mathbf{x}}]\mathbf{y}\right\} \\ &= \left[\mathrm{diag}\left\{\left[\frac{\tilde{a}_1}{\tilde{b}_1},\ldots,\frac{\tilde{a}_M}{\tilde{b}_M}\right]\right\} + \frac{\tilde{c}}{\tilde{d}}\mathbf{G}^H[\tilde{\mathbf{x}}]\mathbf{G}[\tilde{\mathbf{x}}]\right]\tilde{\mathbf{m}}_\mu - \frac{\tilde{c}}{\tilde{d}}\mathbf{G}^H[\tilde{\mathbf{x}}]\mathbf{y}. \end{aligned} \tag{60}$$

Setting (60) to zero leads to (40).

Updating $(\tilde{a}_j, \tilde{b}_j)$ given $\tilde{\mathbf{m}}_\mu, \tilde{\mathbf{\Sigma}}_\mu, \tilde{c}, \tilde{d}$ and $\tilde{\mathbf{x}}_d$

Gathering the terms in $\mathbb{F}$ that are related to $\tilde{a}_j$, we have

$$\begin{aligned}\frac{\partial \mathbb{F}}{\partial \tilde{a}_j} &= \frac{\partial}{\partial \tilde{a}_j}\left\{\mathrm{Tr}\left\{\mathrm{diag}\left\{\left[\frac{\tilde{a}_1}{\tilde{b}_1},\ldots,\frac{\tilde{a}_M}{\tilde{b}_M}\right]\right\}\left(\tilde{\mathbf{m}}_\mu\tilde{\mathbf{m}}_\mu^H + \tilde{\mathbf{\Sigma}}_\mu\right)\right\} - \Psi(\tilde{a}_j)\right. \\ &\quad + \left[\tilde{a}_j\log\tilde{b}_j + (\tilde{a}_j - 1)\left[\Psi(\tilde{a}_j) - \log\tilde{b}_j\right] - \tilde{a}_j - \log\Gamma(\tilde{a}_j)\right] \\ &\quad \left. - \left[(a_j - 1)\left[\Psi(\tilde{a}_j) - \log\tilde{b}_j\right] - b_j\tilde{a}_j/\tilde{b}_j\right]\right\} \\ &= \frac{1}{\tilde{b}_j}\left[|[\tilde{\mathbf{m}}_\mu]_j|^2 + [\tilde{\mathbf{\Sigma}}_\mu]_{j,j}\right] - \Psi'(\tilde{a}_j) + \log\tilde{b}_j + (\tilde{a}_j - 1)\Psi'(\tilde{a}_j) \\ &\quad + \Psi(\tilde{a}_j) - \log\tilde{b}_j - 1 - \Psi(\tilde{a}_j) - (a_j - 1)\Psi'(\tilde{a}_j) + \frac{b_j}{\tilde{b}_j} \\ &= (\tilde{a}_j - a_j - 1)\Psi'(\tilde{a}_j) - 1 + \frac{1}{\tilde{b}_j}\left[|[\tilde{\mathbf{m}}_\mu]_j|^2 + [\tilde{\mathbf{\Sigma}}_\mu]_{j,j} + b_j\right]. \end{aligned} \tag{61}$$

Similarly,

$$
\begin{aligned}
\frac{\partial \mathbb{F}}{\partial \tilde{b}_j} &= \frac{\partial}{\partial \tilde{b}_j}\Big\{ \mathrm{Tr}\left\{ \mathrm{diag}\left\{ \left[\frac{\tilde{a}_1}{\tilde{b}_1}, \ldots, \frac{\tilde{a}_M}{\tilde{b}_M} \right] \right\} \left(\tilde{\mathbf{m}}_\mu \tilde{\mathbf{m}}_\mu^H + \tilde{\boldsymbol{\Sigma}}_\mu \right) \right\} + \log \tilde{b}_j \\
&\quad + \left[\tilde{a}_j \log \tilde{b}_j - (\tilde{a}_j - 1) \log \tilde{b}_j \right] - \left[(a_j - 1)(-\log \tilde{b}_j) - b_j \tilde{a}_j / \tilde{b}_j \right] \Big\} \\
&= -\frac{\tilde{a}_j}{\tilde{b}_j^2} \left[|[\tilde{\mathbf{m}}_\mu]_j|^2 + [\tilde{\boldsymbol{\Sigma}}_\mu]_{j,j} \right] + \frac{1}{\tilde{b}_j} + \frac{\tilde{a}_j}{\tilde{b}_j} - \frac{\tilde{a}_j - 1}{\tilde{b}_j} + \frac{a_j - 1}{\tilde{b}_j} - \frac{b_j \tilde{a}_j}{\tilde{b}_j^2} \\
&= \frac{a_j + 1}{\tilde{b}_j} - \frac{\tilde{a}_j}{\tilde{b}_j^2} \left[|[\tilde{\mathbf{m}}_\mu]_j|^2 + [\tilde{\boldsymbol{\Sigma}}_\mu]_{j,j} + b_j \right]. \qquad (62)
\end{aligned}
$$

Setting both (61) and (62) to zero and solving the simultaneous equations, we obtain (41) and (42).

Updating $(\tilde{c}, \tilde{d})$ given $\tilde{\mathbf{m}}_\mu, \tilde{\boldsymbol{\Sigma}}_\mu, \tilde{a}_j, \tilde{b}_j$ and $\tilde{\mathbf{x}}_d$

Following the procedure in updating other parameters, we compute

$$
\begin{aligned}
\frac{\partial \mathbb{F}}{\partial \tilde{c}} &= \frac{\partial}{\partial \tilde{c}}\Big\{ (\tilde{c} - 1)\, \Psi(\tilde{c}) - \tilde{c} - \log \Gamma(\tilde{c}) - (c - 1)\Psi(\tilde{c}) + \frac{\tilde{c} d}{\tilde{d}} - N \Psi(\tilde{c}) \\
&\quad + \frac{\tilde{c}}{\tilde{d}} \Big[\mathbf{y}^H \mathbf{y} - 2\mathrm{Re}\left\{ \mathbf{y}^H \mathbf{G}[\tilde{\mathbf{x}}] \tilde{\mathbf{m}}_\mu \right\} + \mathrm{Tr}\left\{ \mathbf{G}^H[\tilde{\mathbf{x}}] \mathbf{G}[\tilde{\mathbf{x}}] \left(\tilde{\mathbf{m}}_\mu \tilde{\mathbf{m}}_\mu^H + \tilde{\boldsymbol{\Sigma}}_\mu \right) \right\} \Big] \Big\} \\
&= (\tilde{c} - 1)\Psi'(\tilde{c}) + \Psi(\tilde{c}) - 1 - \Psi(\tilde{c}) - (c - 1)\Psi'(\tilde{c}) + \frac{d}{\tilde{d}} - N \Psi'(\tilde{c}) \\
&\quad + \frac{1}{\tilde{d}} \Big[\mathbf{y}^H \mathbf{y} - 2\mathrm{Re}\left\{ \mathbf{y}^H \mathbf{G}[\tilde{\mathbf{x}}] \tilde{\mathbf{m}}_\mu \right\} + \mathrm{Tr}\left\{ \mathbf{G}^H[\tilde{\mathbf{x}}] \mathbf{G}[\tilde{\mathbf{x}}] (\tilde{\mathbf{m}}_\mu \tilde{\mathbf{m}}_\mu^H + \tilde{\boldsymbol{\Sigma}}_\mu) \right\} \Big] \\
&= (\tilde{c} - c - N)\Psi'(\tilde{c}) - 1 + \frac{1}{\tilde{d}} \Big[d + \mathbf{y}^H \mathbf{y} - 2\mathrm{Re}\left\{ \mathbf{y}^H \mathbf{G}[\tilde{\mathbf{x}}] \tilde{\mathbf{m}}_\mu \right\} \\
&\quad + \mathrm{Tr}\left\{ \mathbf{G}^H[\tilde{\mathbf{x}}] \mathbf{G}[\tilde{\mathbf{x}}] (\tilde{\mathbf{m}}_\mu \tilde{\mathbf{m}}_\mu^H + \tilde{\boldsymbol{\Sigma}}_\mu) \right\} \Big], \qquad (63)
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{\partial \mathbb{F}}{\partial \tilde{d}} &= \frac{\partial}{\partial \tilde{d}}\Big\{ c \log \tilde{d} + \frac{\tilde{c} d}{\tilde{d}} + N \log \tilde{d} + \frac{\tilde{c}}{\tilde{d}} \Big[\mathbf{y}^H \mathbf{y} - 2\mathrm{Re}\left\{ \mathbf{y}^H \mathbf{G}[\tilde{\mathbf{x}}] \tilde{\mathbf{m}}_\mu \right\} \\
&\quad + \mathrm{Tr}\left\{ \mathbf{G}^H[\tilde{\mathbf{x}}] \mathbf{G}[\tilde{\mathbf{x}}] (\tilde{\mathbf{m}}_\mu \tilde{\mathbf{m}}_\mu^H + \tilde{\boldsymbol{\Sigma}}_\mu) \right\} \Big] \} \\
&= -\frac{\tilde{c}}{\tilde{d}^2} [d + \mathbf{y}^H \mathbf{y} - 2\mathrm{Re}\{ \mathbf{y}^H \mathbf{G}[\tilde{\mathbf{x}}] \tilde{\mathbf{m}}_\mu \} + \frac{c + N}{\tilde{d}} + \mathrm{Tr}\{ \mathbf{G}^H[\tilde{\mathbf{x}}] \mathbf{G}[\tilde{\mathbf{x}}] (\tilde{\mathbf{m}}_\mu \tilde{\mathbf{m}}_\mu^H + \tilde{\boldsymbol{\Sigma}}_\mu) \}] (64)
\end{aligned}
$$

Setting both (63) and (64) to zero and solving the two simultaneous equations, we obtain (43) and (44).

REFERENCES

[1] Y.-S. Choi, P. Voltz, and F. Cassara, "On channel estimation and detection for multicarrier signals in fast and selective Rayleigh fading channels," *IEEE Transactions on Communications*, vol. 49, no. 8, pp. 1375–1387, Aug. 2001.

[2] X. Cai and G. Giannakis, "Bounding performance and suppressing intercarrier interference in wireless mobile OFDM," *IEEE Transactions on Communications*, vol. 51, no. 12, pp. 2047–2056, Dec. 2003.

[3] C. Athaudage and A. Jayalath, "Enhanced MMSE channel estimation using timing error statistics for wireless OFDM systems," *IEEE Transactions on Broadcasting*, vol. 50, no. 4, pp. 369–376, Dec. 2004.

[4] R. Negi and J. Cioffi, "Pilot tone selection for channel estimation in a mobile OFDM system," *IEEE Transactions on Consumer Electronics*, vol. 44, no. 3, pp. 1122–1128, Aug. 1998.

[5] Z. Tang, R. Cannizzaro, G. Leus, and P. Banelli, "Pilot-assisted time-varying channel estimation for OFDM systems," *IEEE Transactions on Signal Processing*, vol. 55, no. 5, pp. 2226–2238, May 2007.

[6] T. Al-Naffouri, K. Islam, N. Al-Dhahir, and S. Lu, "A model reduction approach for OFDM channel estimation under high mobility conditions," *IEEE Transactions on Signal Processing*, vol. 58, no. 4, pp. 2181–2193, Apr. 2010.

[7] G. Wang, F. Gao, W. Chen, and C. Tellambura, "Channel estimation and training design for two-way relay networks in time-selective fading environments," *IEEE Transactions on Wireless Communications*, vol. 10, no. 8, pp. 2681–2691, Aug. 2011.

[8] L. He, Y.-C. Wu, S. Ma, T.-S. Ng, and H. Vincent Poor, "Superimposed training based channel estimation and data detection for OFDM amplify-and-forward cooperative systems under high mobility," *to appear in IEEE Transactions on Signal Processing*.

[9] F. Gao, T. Cui, and A. Nallanathan, "On channel estimation and optimal training design for amplify and forward relay networks," *IEEE Transactions on Wireless Communications*, vol. 7, no. 5, pp. 1907–1916, May 2008.

[10] C. Bishop and M. Tipping, "Variational relevance vector machines," in *the 16th Conference on Uncertainty in Artifcial Intelligence*, 2000, pp. 46–53.

[11] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. on Scientific Computing*, no. 1, pp. 129–159.

[12] C. Berger, Z. Wang, J. Huang, and S. Zhou, "Application of compressive sensing to sparse channel estimation," *IEEE Communications Magazine*, vol. 48, no. 11, pp. 164–174, Nov. 2010.

[13] C. Berger, S. Zhou, J. Preisig, and P. Willett, "Sparse channel estimation for multicarrier underwater acoustic communication: from subspace methods to compressed sensing," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1708–1721, Mar. 2010.

[14] G. Taubock, F. Hlawatsch, D. Eiwen, and H. Rauhut, "Compressive estimation of doubly selective channels in multicarrier systems: Leakage effects and sparsity-enhancing processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 255–271, Apr. 2010.

[15] D. Wipf and B. Rao, "Sparse bayesian learning for basis selection," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2153–2164, Aug. 2004.

[16] B. J. Frey, *Graphical models for machine learning and digital communication*. MIT Press, 1998.

[17] S. Haykin *et al.*, *New directions in statistical signal processing: from systems to brain*. MIT Press, 2005.

[18] D. D. Lin and T. J. Lim, "The variational inference approach to joint data detection and phase noise estimation in OFDM," *IEEE Transactions on Signal Processing*, vol. 55, no. 5, pp. 1862–1874, May 2007.

[19] C. M. Bishop, *Pattern Recognition and Machine Learning*, 1st ed. Springer, 2006.

[20] K. Teo and S. Ohno, “Pilot-aided channel estimation and viterbi equalization for OFDM over doubly-selective channel,” in *IEEE Global Telecommunications Conference*, Dec. 2006, pp. 1–5.

[21] M. Tipping, “Sparse bayesian learning and the relevance vector machine,” *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, Sept. 2001.

[22] W. C. Jakes and D. C. Cox, *Microwave Mobile Communications*. Wiley-IEEE Press, 1994.

[23] L. He, S. Ma, Y.-C. Wu, and T.-S. Ng, “Joint channel estimation and data detection for OFDM systems over doubly selective channels,” in *International Symposium on Personal, Indoor and Mobile Radio Communications*, Sept. 2009, pp. 446–450.

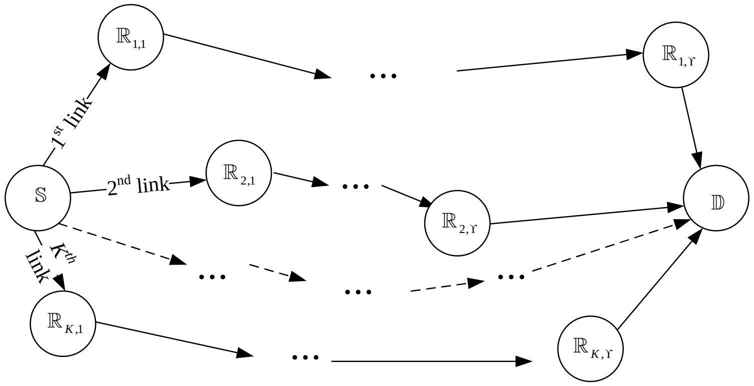

Fig. 1. Multihop Cooperative Communication System

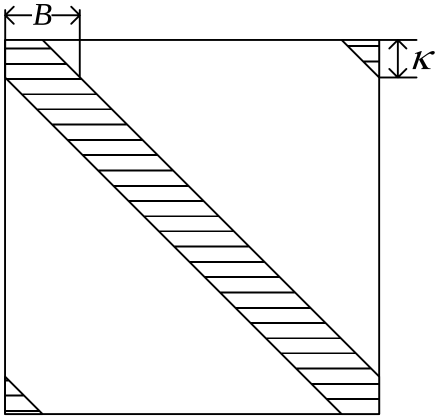

Fig. 2. Banded Matrix Structure Approximation for $\mathbf{D}[\tilde{\mathbf{m}}_{\mu}]$ and $\mathbf{D}[\boldsymbol{\xi}_j]$

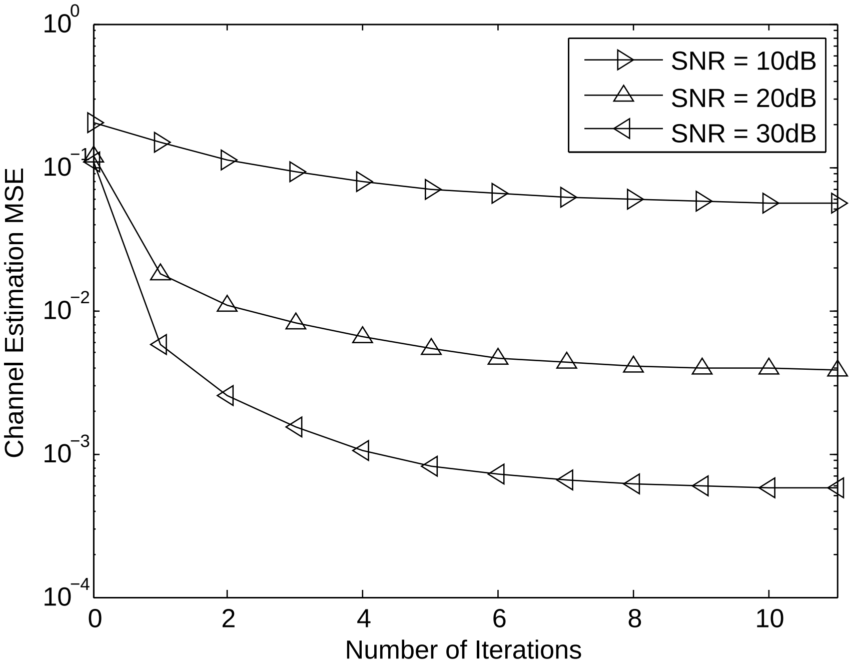

Fig. 3. Convergence of Channel Estimation for Dualhop OFDM System

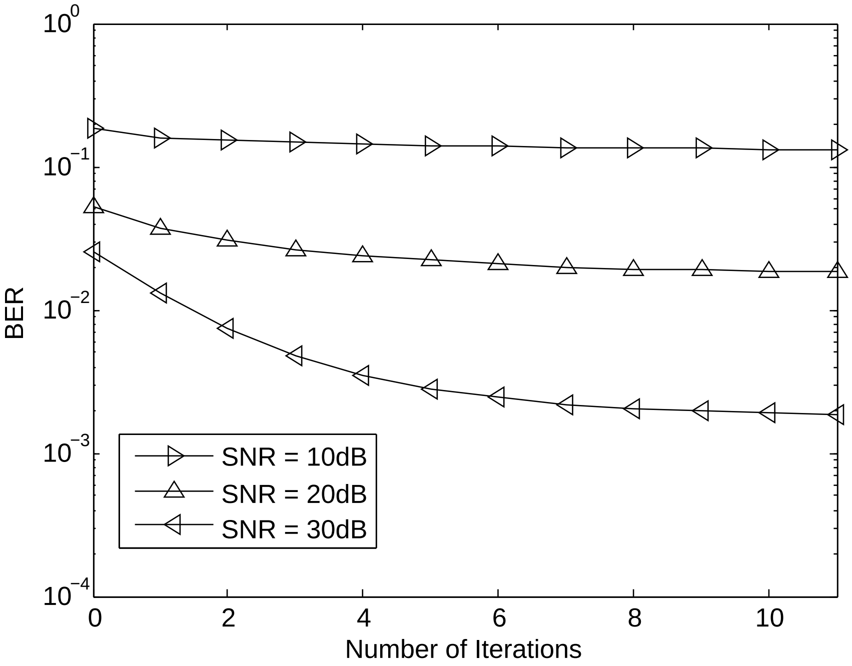

Fig. 4. Convergence of Data Detection for Dualhop OFDM System

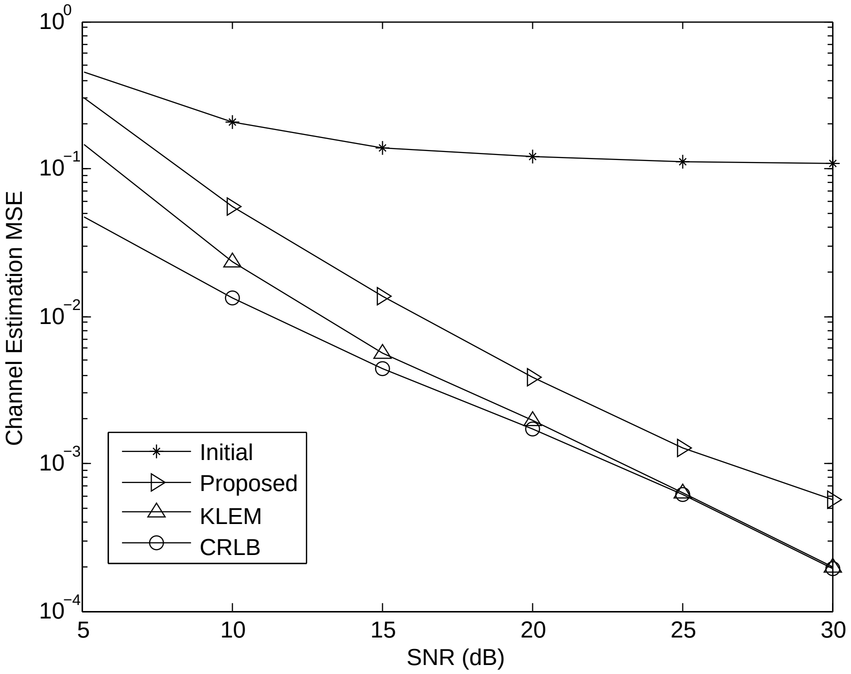

Fig. 5. Performance of Channel Estimation for Dualhop OFDM System

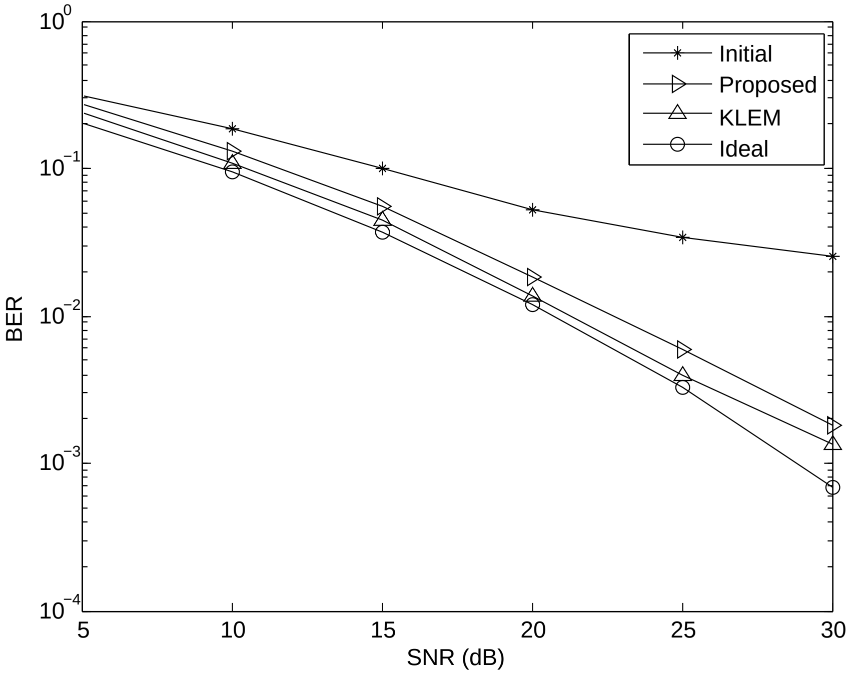

Fig. 6. Performance of Data Detection for Dualhop OFDM System

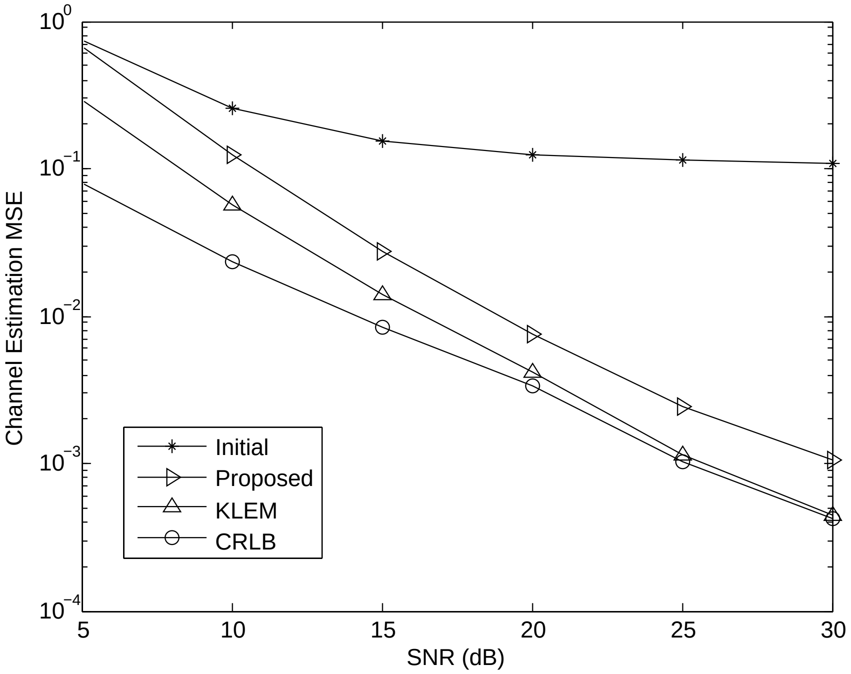

Fig. 7. Performance of Channel Estimation for Three-hop OFDM System

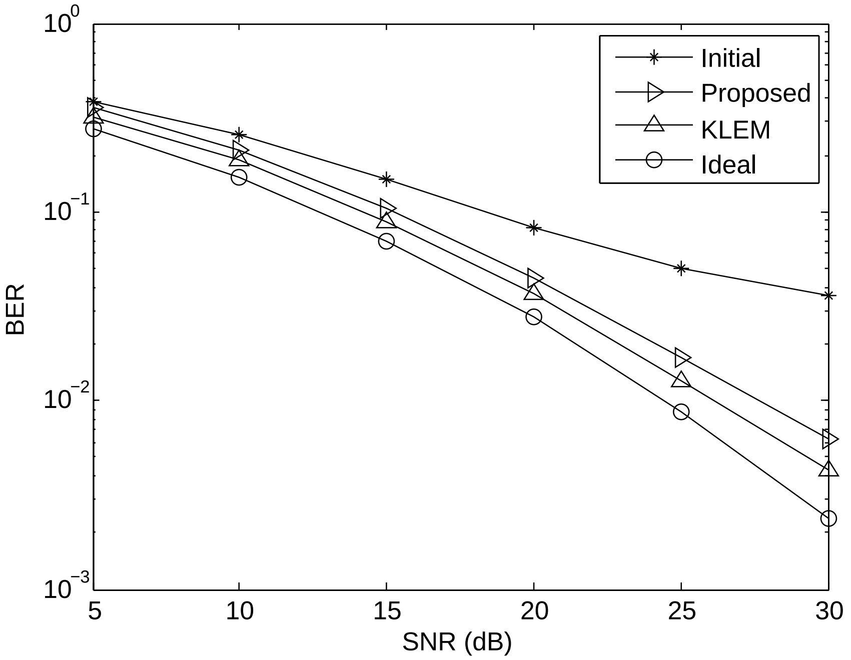

Fig. 8. Performance of Data Detection for Three-hop OFDM System